El Mercado Automovilístico: Un caso para el Análisis Estadístico Multivariante

José Jaén Delgado

1 Introducción

El medio de transporte más utilizado es el coche y por ello un estudio sobre los automóviles puede resultar provechoso para un público considerable: desde los propios usuarios hasta los reguladores políticos. Gracias a que dichos vehículos cuentan con multitud de características mesurables, el Análisis Multivariante resulta una de las técnicas estadísticas más interesantes para obtener información importante.

2 Metodología del Proyecto

A lo largo del presente estudio utilizaremos el lenguaje de programación **Python**, el software estadístico **R** y la plataforma computacional **MATLAB**. Las operaciones siguen una lógica que iremos exponiendo a lo largo del proyecto. Salvo que se especifique lo contrario, la autoría de las diferentes aplicaciones no predifinidas en **MATLAB** de las técnicas estadísticas que emplearemos se atribuye a Baíllo & Grané (2007).

El código para realizar las operaciones se incluirá en un apéndice al final del trabajo.

3 Descripción de los Datos

Los datos se han obtenido de una de las páginas web más conocidas dentro de la comunidad Data Science: ***UCI ML Repository***, de la Universidad de California.

La base de datos con la que trabajaremos cuenta con un total de **199** coches y originalmente fue diseñada con el propósito de desarrollar modelos de Machine Learning para predecir el precio de los automóviles.

Contamos con un total de **11 variables** o rasgos explicativos, de las cuales **9** son **cuantitativas** y el resto **categóricas**. Dentro de estas últimas, trabajeremos con **una variable binaria** y **una multiestado**.

A continuación se presenta una lista detallada de cada variable:

A) Variables Cuantitativas:

- `distancia_ejes`: Distancia entre ruedas delanteras y traseras en centímetros
- `largo`: Medido en centímetros
- `ancho`: Medido en centímetros
- `altura`: Medido en centímetros
- `peso`: En libras
- `motor`: Dimensión del motor
- `caballos`: Potencia del coche en caballos de vapor
- `max_revoluciones`: Máximo de revoluciones por minuto
- `precio`: Cuantía de venta en dólares

B) Variable Binaria:

- `combustible`: Toma valor 1 si es gasolina, en caso de diesel es 0

C) Variable Multiestado:

- `rueda_motriz`: Tracción trasera (valor 1), Tracción delantera (valor 2), 4x4 (valor 3)

Nótese que se ha realizado una selección concreta de las variables del conjunto de datos original, pues en realidad se contaba con un total de 26 variables. Aún así el número de rasgos explicativos es significativo, permitiendo un Análisis Multivariante correcto.

4 Análisis Variables Cuantitativas

En esta sección realizamos un análisis descriptivo multivariante centrándonos en las variables cuantitativas de nuestra base de datos y posteriormente un ejercicio de inferencia estadística. Para lo primero nos serviremos del vector de medias, herramientas visuales y las matrices de covarianzas y correlaciones.

En la diagonal del gráfico de dispersión matricial se puede apreciar la representación de las distribuciones de las diferentes variables.

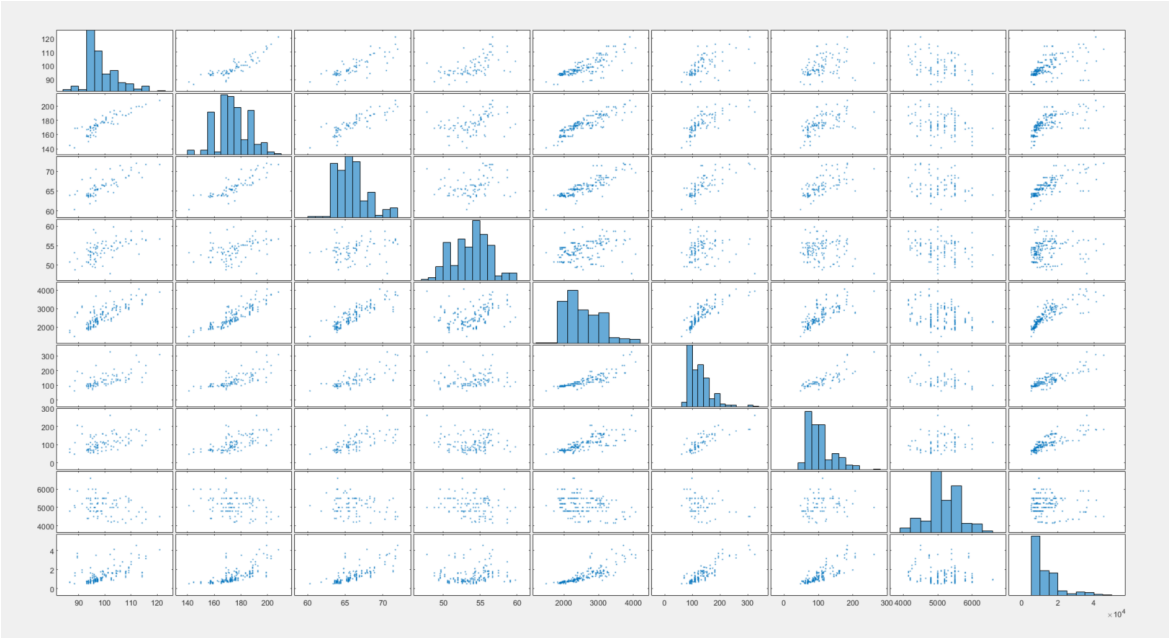

Figure 1: Gráfico de Dispersión Matricial Original

A simple vista las relaciones lineales entre las características de los coches se presentan positivas, si bien en el caso de las revoluciones máximas (ocatava fila) es más bien inexistente. Estos resultados son lógicos y por tanto permiten una interpretación coherente, ya que las variables describen las dimensiones de los coches. Consecuentemente y como ejemplo ilustrativo, a mayor tamaño, peso o potencia (caballos de vapor), se observa un aumento en el precio y resto de características.

Resulta esperable que los coches más grandes o con mayor número de caballos sean a su vez los más caros, largos o pesados. Esto es precisamente la asociación que muestra el gráfico de dispersión matricial de la Figura 1.

Al objeto de facilitar el análisis posterior, intentaremos que las distribuciones de las variables se asemejen a una normal mediante una serie de transformaciones no lineales. Programamos un algoritmo en Python que permita reconocer qué tipo de operación (logaritmo o raíz cuadrada) aumenta la simetría de la distribución de las variables basándonos en el valor de la inclinación o *skewness*. Aprovechamos la mención al lenguaje de programación Python para recordar al lector que en el apéndice del presente trabajo se facilitarán dos enlaces: uno al Jupyter Notebook de Python y otro al repositorio de GitHub donde se podrá encontrar el código de MATLAB y R utilizados.

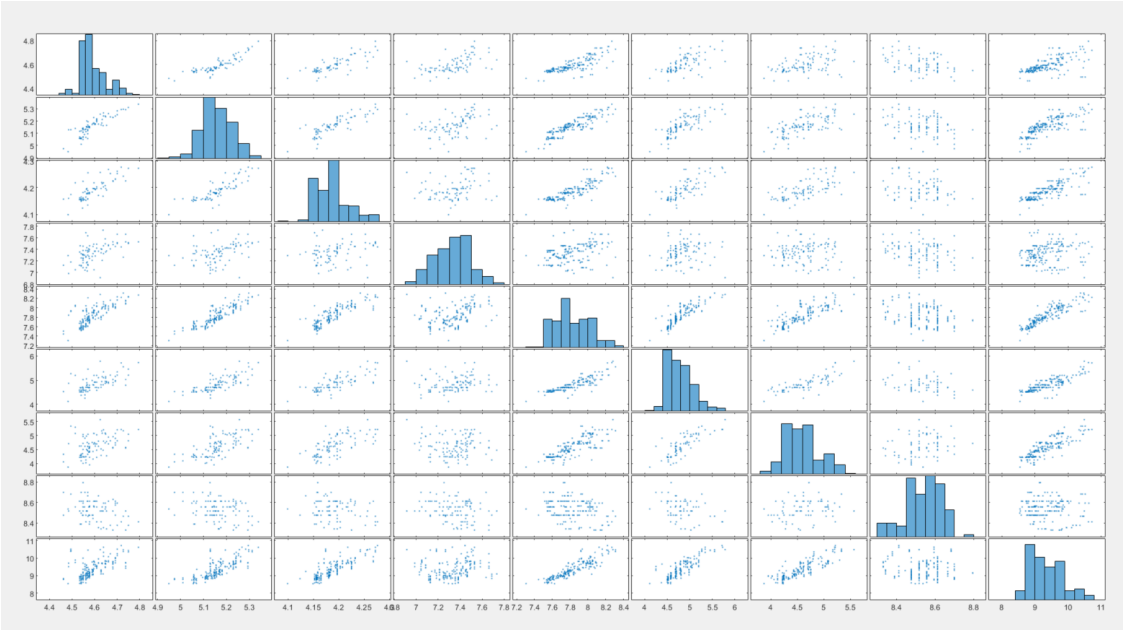

Figure 2: Gráfico de Dispersión Matricial con Variables Transformadas

Pese a no conseguir emular de manera exacta la distribución normal, las operaciones realizadas resultan en una mejora de la simetría de las variables.

Presentamos a continuación un resumen numérico de la medida de centralidad escogida a modo de análisis de cada variable: el vector de medias, así como un estudio descriptivo de las relaciones entre columnas en forma de matriz de covarianzas y matriz de correlaciones.

La primera tabla correspondiente al vector de medias muestra dicha medida tanto en el conjunto de datos originales como en las variables transformadas. Inlcuimos asimismo la diferencia entre ambos para mostrar el cambio.

Table 1: Vector de Medias

Variables	Originales	Nuevas	$(\bar{X}_i - \mathring{X}_i)$	Transformación
distancia_ejes	98.789	4.591	94.198	$\log(X_i)$
largo	174.097	5.157	168.934	$\log(X_i)$
ancho	65.874	4.187	61.687	$\log(X_i)$
altura	53.774	7.331	46.443	$\sqrt{X_i}$
peso	2552.925	7.825	2545.10	$\log(X_i)$
motor	126.667	4.798	121.868	$\log(X_i)$
caballos	103.23	4.578	98.652	$\log(X_i)$
max_revoluciones	5119	8.536	5110.464	$\log(X_i)$
precio	13223.69	9.350	13214.339	$\log(X_i)$

Nota: $\mathring{X}_i$ denota la media de la variable i transformada, $\bar{X}_i$ el caso original

Asimismo, presentamos la tabla correspondiente a la matriz de covarianzas. En la diagonal observamos que las varianzas son muy distintas entre ellas, presentando cada columna original de nuestra matriz de datos una variabilidad muy diferente. Tras aplicar las transformaciones pertinentes para aumentar la simetría de las distribuciones las varianzas se asemejan más entre ellas. Por otro lado, centrándonos en los signos de las covarianzas concluimos que allá donde éste sea positivo es indicativo de que dichas variables exhiben una relación lineal positiva entre ambas, y en otro caso muestran una evolución contraria. Para poder pronunciarnos sobre la intensidad de la relación entre variables será necesario centrarnos en la matriz de correlaciones.

Table 2: Matriz de Covarianzas Originales

	dist_ejes	largo	ancho	altura	peso	motor	caballos	max_re	precio
dist_ejes	37.1								
largo	66.28	153.06							
ancho	10.51	22.39	4.46						
altura	8.84	14.98	1.6	5.99					
peso	2478.1	5670.67	951.42	391.66	270249.79				
motor	145.77	353.98	64.3	7.72	18434.5	1743.31			
caballos	85	269.62	48.78	-8.03	14799.33	1290	1410.29		
max_rev	-1055.38	-1702.17	-249.51	-366.37	-69788.41	-5151.69	1946.77	230901.22	
precio	28370.78	68502.52	12697.47	2635.58	3463767.6	291121.83	242860.94	-389715.97	63659775.11

Table 3: Matriz de Covarianzas Transformadas

	distancia_ejes	largo	ancho	altura	peso	motor	caballos	max_revoluciones	precio
distancia_ejes	0.0036								
largo	0.0037	0.005							
ancho	0.0015	0.0019	0.001						
altura	0.006	0.0058	0.0016	0.0279					
peso	0.0091	0.0125	0.0054	0.0102	0.0389				
motor	0.0101	0.0147	0.0067	0.0061	0.0482	0.0802			
caballos	0.0085	0.0151	0.0069	-0.0033	0.0527	0.0778	0.1142		
max_revoluciones	-0.0021	-0.0019	-7e-04	-0.005	-0.0051	-0.0077	0.0036	0.0089	
precio	0.0191	0.0278	0.0128	0.0149	0.0889	0.1208	0.1433	-0.0054	0.2545

Table 4: Matriz de Correlaciones Originales

	distancia_ejes	largo	ancho	altura	peso	motor	caballos	max_revoluciones	precio
distancia_ejes	1								
largo	0.88	1							
ancho	0.82	0.86	1						
altura	0.59	0.49	0.31	1					
peso	0.78	0.88	0.87	0.31	1				
motor	0.57	0.69	0.73	0.08	0.85	1			
caballos	0.37	0.58	0.62	-0.09	0.76	0.82	1		
max_revoluciones	-0.36	-0.29	-0.25	-0.31	-0.28	-0.26	0.11	1	
precio	0.58	0.69	0.75	0.13	0.84	0.87	0.81	-0.1	1

Table 5: Matriz de Correlaciones Transformadas

	distancia_ejes	largo	ancho	altura	peso	motor	caballos	max_revoluciones	precio
distancia_ejes	1								
largo	0.87	1							
ancho	0.81	0.85	1						
altura	0.6	0.49	0.31	1					
peso	0.77	0.89	0.86	0.31	1				
motor	0.59	0.73	0.75	0.13	0.86	1			
caballos	0.42	0.63	0.65	-0.06	0.79	0.81	1		
max_revoluciones	-0.37	-0.29	-0.25	-0.32	-0.27	-0.29	0.11	1	
precio	0.63	0.77	0.8	0.18	0.89	0.85	0.84	-0.11	1

La diferencia entre las matrices de correlaciones es mínima. Ambas muestran relaciones lineales fuertes (la mayoría superiores a 0.6) y positivas entre variables, a excepción de la variable ya indicada (las revoluciones). Ésta última apenas está correlada con el resto y además presenta un signo diferente.

Nos centramos ahora en la obtención de medidas escalares de dispersión, a saber, variación total, variación generalizada y la interdependencia lineal o η^2.

La variación total se expresa como la traza de la matriz de covarianzas o la suma de los autovalores de dicha matriz $\mathbf{S}$.

La varianza generalizada es el determinante de $\mathbf{S}$ o el producto de los autovalores de $\mathbf{S}$, y por último, la medida de interdependencia lineal definida como η^2 se expresa como:

$$\eta^2 = 1 - \det(\mathbf{R})$$

Donde:

$$\mathbf{S} = \tfrac{1}{n}\mathbf{X}'\mathbf{H}\mathbf{X}$$
$$\mathbf{R} = \mathbf{D}_{\mathrm{s}}^{-1}\mathbf{S}\mathbf{D}_{\mathrm{s}}^{-1}$$
$$\mathbf{D}_{\mathrm{s}} = \mathrm{diag}((\sqrt{S_{11}}, \sqrt{S_{22}}, \ldots \sqrt{S_{pp}})')$$

Table 6: Medidas de Dispersión Multivariante

Medidas	Originales	Nuevas
Varianza Generalizada	3.5339e+25	7.4029e-21
Variación Total	6.3842e+07	0.5315
η^2	1	1

En ambos conjuntos de datos existen relaciones lineales entre las variables, dado que $\eta^2 \approx 1$.

Verificamos que tanto la Variación Generalizada como la Total son más reducidos en el caso de las variables transformadas, y por ello más convenientes para el resto de operaciones.

Concluimos que en nuestro conjunto de datos existe relación lineal entre las variables, mayoritariamente positiva y fuerte. Las variables transformadas parecen más convenientes para un futuro análisis debido a su semejanza con la distribución normal y los valores de la variación total y varianza generalizada.

5 Análisis Variable Binaria

En esta sección realizamos un breve análisis del conjunto de datos transformados basado en el valor de la variable binaria `combustible`.

Recordamos que ésta toma el valor 1 si el coche funciona mediante gasolina, y 0 en caso de que se utilice diesel. Antes de proseguir con ninguna operación o gráfico es necesario aclarar que nos encontramos ante una **base de datos desequilibrada** o *imbalanced*, dado que los coches con gasolina suponen el 90% mientras que los de diesel un 10%.

Sería posible mitigar los efectos negativos del desequilibrio entre grupos sirviéndonos de algoritmos de remuestreo tales como el **Random Oversampling** o el **Random Undersampling**. Asimismo, podría emplearse técnicas más avanzadas como el **SMOTE-ENN** (Synthetic Minority Oversampling TEchnique - Edited Nearest Neighbor), pero no aplicaremos estas técnicas dado que escapan del objetivo de este proyecto.

Aportamos gráfico de dispersión matricial desagregado por los grupos ya mencionados:

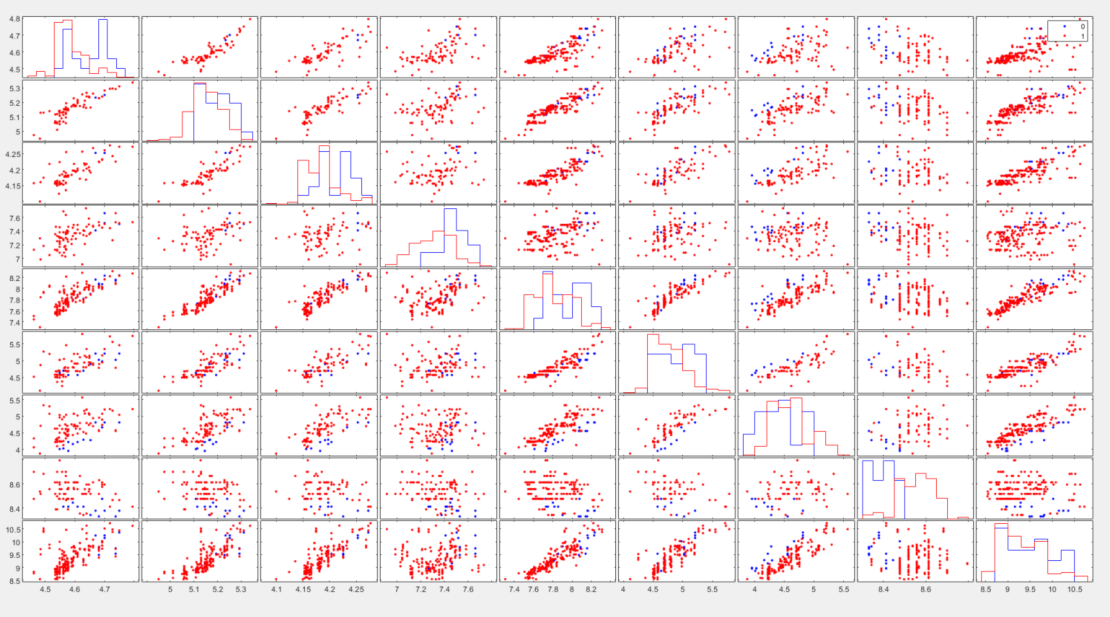

Figure 3: Gráfico de Dispersión Matricial con Variables Transformadas

A excepción del precio, los histogramas varían considerablemente entre grupos, lo que supone que las distribuciones de las variable de los coches a gasolina son diferentes de los automóviles con diesel. Por tanto, es esperable que las características

técnicas varíen entre un tipo de vehículos y otro. En cuanto a las relaciones lineales, prácticamente todas siguen una misma tendencia.

Acudiendo a la inferencia estadística, aplicaremos un contraste de hipótesis sobre la igualdad de las medias de ambos grupos. De esta manera podremos estudiar si existen diferencias entre dichas medidas de centralidad. Para ello utilizaremos el estadístico T^2 de Hotelling, asumiendo que las filas de nuestra matriz de datos provienen de leyes normales $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ independientes, por lo que las transformaciones de la sección anterior nos acercan a esta suposición.

$$\mathrm{H}_0 : \boldsymbol{\mu_X} = \boldsymbol{\mu_Y}$$
$$\mathrm{H}_1 : \boldsymbol{\mu_X} \neq \boldsymbol{\mu_Y}$$

El estadístico se expresa como:

$$\frac{n_x n_y}{n_x + n_y} \left(\bar{\boldsymbol{x}} - \bar{\boldsymbol{y}}\right)' \mathbf{S}_p^{-1} \left(\bar{\boldsymbol{x}} - \bar{\boldsymbol{y}}\right) \sim T^2(p, n_x + n_y - 2)$$

Donde:

$$\mathbf{S}_p = \tfrac{1}{n_x+n_y}(n_x \mathbf{S}_x + n_y \mathbf{S}_y)$$

Debido a la falta de tabulación de dicha distribución, será necesario transformarla a una F de Fisher aprovechando que se cumple:

$$\tfrac{n-p+1}{np}\, T^2(p, n) = F(p, n - p + 1)$$

Consecuentemente trabajaremos con:

$$\frac{n_x + n_y - p - 1}{(n_x + n_y - 2)p} \frac{n_x n_y}{n_x + n_y} \left(\bar{\boldsymbol{x}} - \bar{\boldsymbol{y}}\right)' \mathbf{S}_p^{-1} \left(\bar{\boldsymbol{x}} - \bar{\boldsymbol{y}}\right) \sim F(p, n_x + n_y - p - 1)$$

Concretamente los grados de libertad son $F(9, 191)$ y el cuantil 0.95 o valor crítico es 1.9292. Aplicando las relaciones previamente mostradas obtenemos un estadístico igual a 18.6439, claramente superior al valor crítico. Por tanto, los datos han arrojado suficiente evidencia en contra de la hipótesis nula de igualdad de medias entre grupos. Rechazamos con un nivel de confianza del 95% que los coches a gasolina y los de diesel presenten un vector de medias igual.

6 Análisis Variable Categórica Multiestado

De manera análoga a la sección anterior, nos centramos en la variable multiestado `rueda_motriz`.

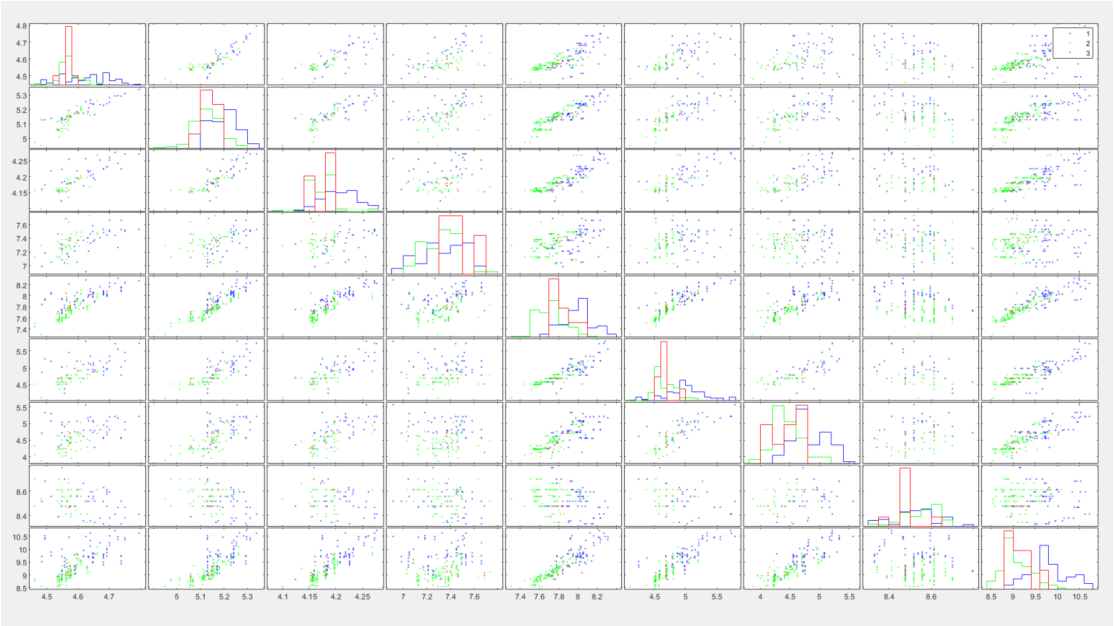

Figure 4: Gráfico de Dispersión Matricial con Variables Transformadas

Si bien ocurre como antes y las relaciones lineales entre los grupos de la variable multiestado son similares, es decir, los coches con tracción trasera, delantera y 4x4 muestran relaciones prácticamente análogas, los histogramas en la diagonal del gráfico de dispersión matricial no son del todo iguales. Las características técnicas de los 4x4 tienden a concentrarse en valores muy concretos, mientras que en el resto de coches se aprecia mayor variabilidad. Una vez más nos encontramos con el problema del desequilibrio entre grupos. Hay un total de 75 vehículos con tracción trasera, 118 con tracción delantera y los restantes 8 son 4x4. Aunque sea verdad que los dos primeros grupos no muestran una gran disparidad, los últimos están claramente infrarepresentados, de ahí que los valores concentrados que mencionamos anteriormente encuentren su explicación en la poca cantidad de datos disponibles.

Realizaremos un contraste de vector de medias entre grupos mediante el estadístico Lambda de Wilks, que como ocurría previamente aproximaremos mediante una F de Fisher.

Contrastamos:

$$\mathrm{H}_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}_3$$
$$\mathrm{H}_1 : \text{Algún } \boldsymbol{\mu}_i \neq \boldsymbol{\mu}_j$$

El estadístico se expresa como:

$$\Lambda = \frac{|\mathbf{W}|}{|\mathbf{W} + \mathbf{B}|} \sim \Lambda(p, n - g, g - 1)$$

Donde $\mathbf{W}$ es la dispersión dentro de los grupos, $\mathbf{B}$ representa la dispersión entre grupos y g las matrices de datos. Matemáticamente:

$$\mathbf{B} = \sum_{i=1}^{g} n_i \left(\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{x}}\right) \left(\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{x}}\right)'$$

$$\mathbf{W} = \sum_{i=1}^{g} \sum_{k=1} \left(x_{ik} - \bar{\boldsymbol{x}}_i\right) \left(x_{ik} - \bar{\boldsymbol{x}}_i\right)'$$

Mediante la aproximación asintótica de Rao:

$$\frac{1 - \Lambda^{1/\beta}}{\Lambda^{1/\beta}} \frac{\alpha\beta - 2\gamma}{pb} \sim F(pb, \alpha\beta - 2\gamma)$$

Es necesario suponer que las g matrices de datos provinen de distribuciones normales multivariantes independientes, de ahí que haya resultado útil transformar las variables durante el análisis de las variables cuantitativas.

El valor crítico de la distribución F de Fisher es 1.6311, y calculando el estadístico obtenemos 21.4316, por lo que se concluye que los datos aportan evidencia en contra de la hipótesis nula. Con un nivel de confianza del 95% rechazamos que el vector de medias de cada grupo sea igual.

7 Reducción de Dimensionalidad

Antes de dar paso a la aplicación de las técnicas estadísticas procedemos a exponer brevemente el objetivo que se persigue con las mismas, explicando su utilidad.

Actualmente, el ***Big Data*** o volumen masivo de datos es uno de los temas dominantes en el mundo de la Analítica. Esto se explica en buena parte debido a los grandes avances de las tecnologías de ***Cloud Computing*** y la aparición de aplicaciones capaces de procesar bases de datos complejas y de considerable tamaño. Por ello, se precisan de métodos que faciliten el análisis de dicha información.

Los algoritmos de ***Machine Learning*** y ***Deep Learning*** cuando son aplicados a grandes bases de datos presentan usos de memoria y espacios de tiempo de entrenamiento distantes de ser calificados de intrascendentes. Igualmente, cualquier modelo que utilice todas las variables explicativas para tareas de predicción o clasificación puede 'sobreajustarse' a los datos disponibles (*overfitting*). El fin de los modelos no es tanto explicar de forma exhaustiva una base de datos específica, sino generalizar de forma correcta a nuevas observaciones.

Consecuentemente, contar con muchas variables explicativas da lugar a la 'maldición de la dimensionalidad' o ***curse of dimensionality***: si bien es deseable que el investigador tenga disponible información suficiente, también es cierto que la complejidad computacional y la precisión de los modelos se resienten.

De forma meramente informativa, entre las técnicas de reducción de la dimensionalidad detacan la ***Regularización L1*** o ***Regresión LASSO*** propuesta por Tibshirani (1996), la cual asigna valores nulos a coeficientes de las variables más prescindibles. En la misma línea se encuentra el uso de ***Autoencoders***, idea originalmente desarrollada por Ballard (1987), consistente en Redes Neuronales Artificiales que a partir de los propios inputs descubren estructuras dentro de los datos que permiten un *feature learning* eficiente. Se obtiene una codificación de los datos que posteriormente se valida con las variables originales.

En este proyecto ponemos en práctica las ideas expresadas, concretamente mediante el ***Análisis de Componentes Principales*** (PCA por sus siglas en inglés) y el ***Multidimensional Scaling*** (MDS). Conviene señalar como limitación que tanto el número de observaciones como el de variables, 199 y 11 respectivamente, no presentan un serio problema en términos de logística de modelización o análisis. No obstante, como ejemplo ilustrativo de la utilidad de ambas técnicas y de la posibilidad de descubrir información valiosa resultan del todo convenientes.

8 Análisis de Componentes Principales

La siguiente sección está dedicada al empleo de PCA en nuestro conjunto de datos, técnica desarrollada por Pearson (1921) y Hotelling (1933). El objetivo principal será describir la información contenida en la **matriz centrada** de **datos cuantitativos** mediante un conjunto de variables menor que el de variables originales. Dicho fin se conseguirá aprovechando la correlación existente entre las columnas de la matriz previamente mencionada.

Por el Teorema de la Dimensión, si el rango de la matriz de covarianzas, r, definido como $r = rang(\mathbf{S})$ es menor o igual que el número de variables (p) la diferencia entre éste y el rango representa las p - r variables que son combinaciones lineales de otras (y por tanto las podremos descartar sin pérdida de generalidad). De forma análoga se puede emplear la matriz de correlaciones $\mathbf{R}$ para llegar a la misma conclusión. La elección de $\mathbf{S}$ o $\mathbf{R}$ como base para realizar PCA dependerá de la varianza de las variables cuantitativas, optando por $\mathbf{R}$ si los valores difieren sustancialmente.

En nuestro caso los valores de las varianzas son parecidos únicamente para las variables medidas en centímetros, por lo que las conclusiones principales las obtendremos en base a $\mathbf{R}$. Modificamos la función `comp2()` de Baíllo & Grané (2007) para dividir los elementos de los Componentes Principales por su desviación típica.

Dado que la relación entre las variables es tan importante para el uso correcto de PCA, calculamos una medida escalar de interdependencia lineal, η^2 o coeficiente de intensidad de relaciones lineales entre los datos cuantitativos, definido como:

$$\eta^2 = 1 - \det(\mathbf{R})$$

Para nuestros datos η^2 es muy cercano a 1 (ausencia de incorrelación) por lo que no es esperable que muchos Componentes Principales sean necesarios para describir toda la información importante de la matriz.

Definimos los Componentes Principales $\tilde{\mathbf{Y}}$ como una combinación obtenida a partir de las variables originales:

$$\tilde{\mathbf{Y}} = \mathbf{X}\mathbf{D}_s^{-1}\tilde{\mathbf{T}}, \quad \mathbf{D}_s^{-1} \coloneqq \text{diag}(\sqrt{S_{11}}, \ldots, \sqrt{S_{pp}})$$

Donde $\tilde{\mathbf{T}}$ son los autovectores de la matriz de correlaciones $\mathbf{R}$ de los datos cuantitativos. Aprovechamos que $\mathbf{cR} \geq 0$ ($\forall \mathbf{c} \neq 0$) y simétrica por lo que su descomposición espectral es:

$$\mathbf{R} = \tilde{\mathbf{T}}\tilde{\mathbf{\Lambda}}\tilde{\mathbf{T}}'$$

Para determinar el número de Componentes Principales finales utilizaremos el Criterio de Kaisser modificado o corrección de Jollife. No incluimos los componentes cuyos autovalores sean menores que 0.7, ya que se ha comprobado que cuando $p \leq 20$ (como en nuestro caso) el criterio de Kaisser tiende a incluir pocos componentes.

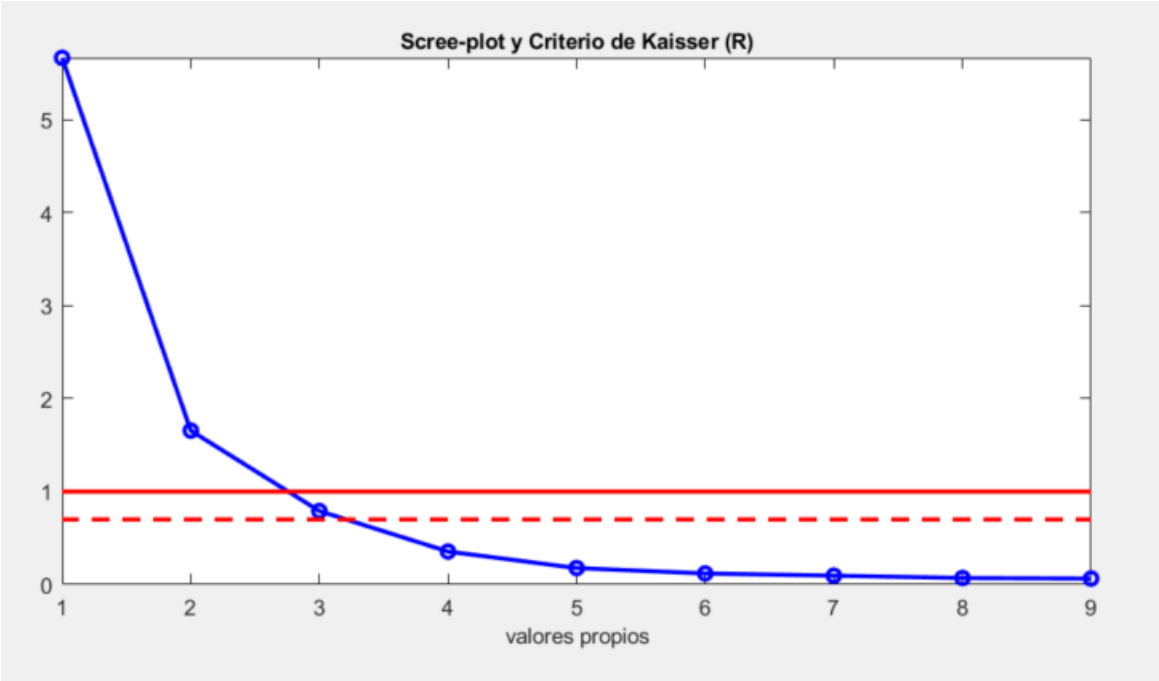

Figure 5: Selección PCA con Criterio Kaisser Modificado

Apreciamos el efecto de la corrección de Jollife (línea discontinua roja) al incluir un tercer componente principal que ha sido excluido por el criterio de Kaisser (línea continua roja). Por tanto, interpretaremos los resultados relevantes empleando tres Componentes Principales.

Obtenemos los valores de cada componente a continuación:

$$
\begin{aligned}
Y_1 &= 0.0584X_1 + 0.0637X_2 + 0.0633X_3 + 0.0259X_4 \\
&\quad + 0.0669X_5 + 0.0607X_6 + 0.0520X_7 - 0.0209X_8 + 0.0605X_9 \\
Y_2 &= -0.0252X_1 - 0.0124X_2 - 0.0024X_3 - 0.0480X_4 \\
&\quad + 0.0032X_5 + 0.0185X_6 + 0.0365X_7 + 0.0372X_8 + 0.0199X_9 \\
Y_3 &= 0.0840X_1 + 0.0799X_2 + 0.0318X_3 + 0.1943X_4 \\
&\quad - 0.0124X_5 - 0.1272X_6 + 0.0124X_7 + 0.3958X_8 - 0.0140X_9
\end{aligned}
$$

La representación gráfica de dichos componentes quedaría de la siguiente forma:

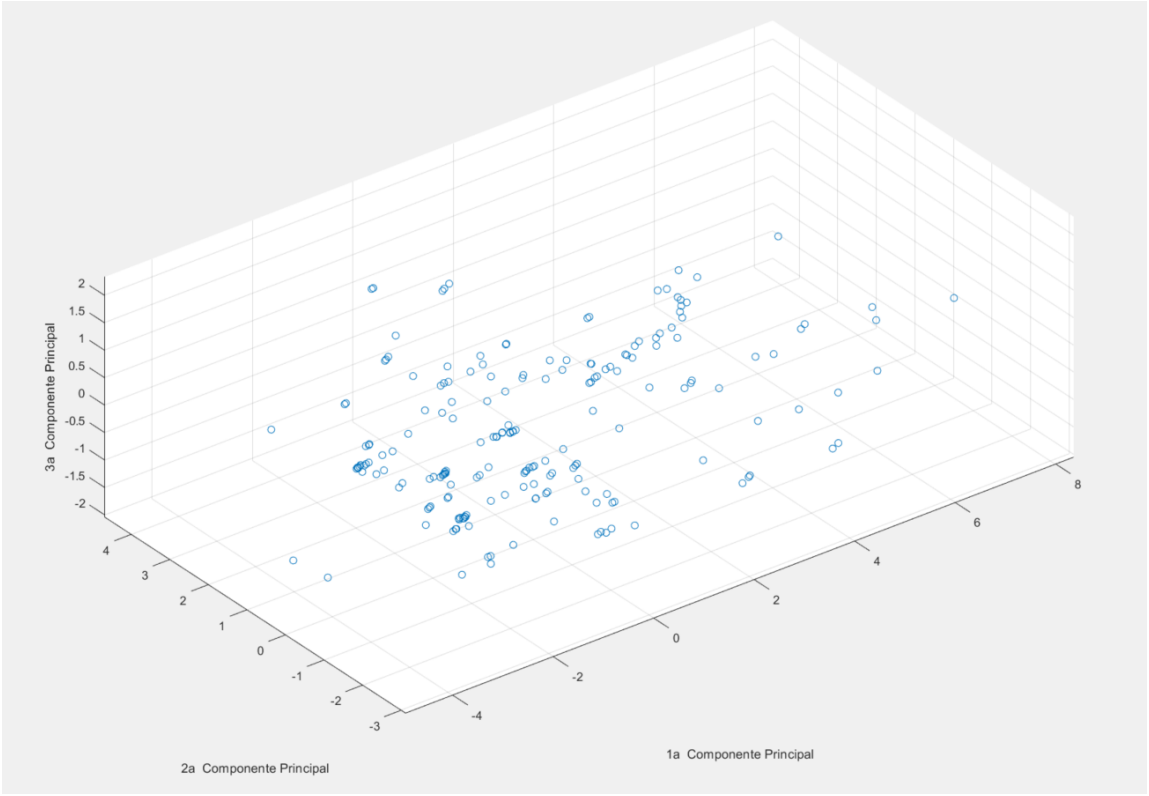

Figure 6: PCA a partir de $\mathbf{R}$ (90.16 %)

Los componentes Y_1, Y_2, Y_3 explican el 90.16% de la variabilidad de la matriz de datos, por lo que la modificación de Jollife coincide con el valor usual del criterio de Porcentaje Explicado (al menos 90%), escenario que no se daría si únicamente explotáramos la información propuesta por los dos primeros componentes (81.37%) y mucho menos por el primero solamente (62.96%).

Asimismo, una rápida revisión de la Figura 1 revela que los descensos de pendiente son poco significativos a partir del tercer autovalor (sería posible considerar los dos primeros autovalores aunque es una perspectiva muy agresiva bajo nuestro punto de vista). Consecuentemente, dado que el criterio de Porcentaje Explicado y el Scree test de Cattell (visualización de pendientes) coinciden, nuestra decisión de incluir tres Componentes Principales a partir del criterio de Kaisser modificado se ve reforzada.

Procediendo a la interpretación de los Componentes Principales cabe apuntar que no nos consideramos expertos en la materia objeto de estudio y por tanto las asociaciones que presentamos pueden no corresponderse fielmente con la realidad técnica de los

automóviles. Pese a que la subjetividad es ubicua en la Estadística (sirva como ejemplo la selección de un a priori adecuado en los Métodos Bayesianos), es deseable contrastar información con profesionales del área del conocimiento que se está analizando. En nuestro caso hemos consultado funtes de información especializadas en coches.

Es reseñable que en Y_1 todos los coeficientes son positivos (a excepción de X_8 o `max_revoluciones`), por lo que podría considerarse un buen proxy del tamaño del coche. Siendo cierto que los índices de tamaño en PCA requieren signos positivos en todos los coeficientes, hemos de señalar que en el proyecto anterior `max_revoluciones` era una de las variables que presentaba correlación más débil con el resto, y de hecho en valor absoluto es la que menos contribuye a Y_1 (`largo`, `ancho`, `peso` las que más).

Para Y_2 ocurre un caso diferente, los cuatro primeros coeficientes son negativos, coincidentes con las dimensiones del coche medidas en centímetros. La aseguradora británica Zuto (2020) realizó un estudio concluyendo que los coches modernos pesan un 70% más que los antiguos dentro de la misma gama. Además, los avances tecnológicos han permitido mejorar considerablemente la potencia de los automóviles, incrementando los caballos de vapor, las revoluciones máximas y la dimensión del motor. Sobre este último punto, la compañía LeasePlan (2017) apunta que los coches eléctricos presentan un mayor motor, dado que para producir un mayor amparaje se requieren bobinas más grandes. Por tanto, Y_2 puede interpretarse como un índice de modernidad del coche, siendo `altura`, `caballos` y `max_revoluciones` las variables más determinantes (valor absoluto mayor).

Finalmente, en el caso de Y_3, nuestro último Componente Principal, la tarea interpretativa se vuelve más intricada. Por simplicidad obviaremos las variables con menor valor absoluto. Nos centramos pues en `max_revoluciones`, `altura`, `motor` y `distancia_ejes`. Únicamente `motor` tiene un valor negativo, por lo que Y_3 es candidato a discernir entre los coches de grandes dimensiones aquellos que mejor optimizan el tamaño del motor. Se penaliza asimismo a coches de pequeña dimensión, y aún más a éstos mismos que no aprovechen los recursos para fabricar un motor potente que no ocupe mucho espacio.

De esta manera, mientras un coche registre datos cuyos valores absolutos sean de considerable cuantía, lo interpretaremos como un automóvil de gran 'tamaño' (proxy), y por tanto obtendrá grandes resultados positivos en el eje '1ª PCA' de nuestro *gráfico tridimensional*. Si el vehículo en cuestión sobresale en cuanto a su peso y no destaca tanto por su altura o anchura sino por las características técnicas, es indicativo de que

es un modelo moderno o con tecnología sofisticada (hacia los puntos máximos del eje '2ª PCA'). Por último, el eje '3ª PCA' muestra en sus valores superiores aquellos coches de importante altura, revoluciones máximas y distancia entre los ejes que maximizan la potencia dado un menor motor. En los valores intermedios de dicho eje podremos encontrar coches más pequeños pero eficientes en la creación del motor.

En cuanto a las correlaciones, un breve repaso de las mismas nos puede ayudar a identificar la procedencia de los diferentes componentes principales.

Table 7: Correlaciones con Componentes Principales

Variables	Y_1	Y_2	Y_3
`distancia_ejes`	0.8441	-0.3997	0.1574
`largo`	0.9207	-0.1974	0.1496
`ancho`	0.9151	-0.0376	0.05947
`altura`	0.3749	-0.7629	0.3639
`peso`	0.9669	0.0502	-0.0231
`motor`	0.8772	0.2946	-0.2383
`caballos`	0.7521	0.5793	0.0232
`max_revoluciones`	-0.3028	0.5916	0.7413
`precio`	0.8756	0.3154	-0.0262

El primer componente principal guarda una gran correlación con el peso, largo, ancho, dimensión del motor, precio y distancia entre los ejes, siendo asimismo fuerte con los caballos de vapor. Tiene sentido pues interpretar Y_1 como una aproximación al tamaño del coche, puntualizando una vez más que la correlación con `max_revoluciones` es débil.

Por su parte, Y_2 está correlada fuertemente con la altura, seguido de los caballos, las revoluciones máximas y el peso, rasgos distintivos de la modernidad del coche como habíamos apuntado. Los valores de las correlaciones empiezan a resentirse y son significativamente menores que Y_1, dado que la variabilidad explicada de los datos cuantitativos es menor.

Y_3 presenta correlaciones menores, explicando en menor medida la información que tenemos. Destacan las revoluciones máximas seguidas por la altura, el motor y distancia entre ejes. La interpretación de este componente es compleja y consideramos que puede servir más bien para identificar la sofisticación ingenieril del coche, en tanto en cuanto premia a aquellos vehículos con menor motor.

9 Multidimensional Scaling

Como uno de los inconvenientes de PCA identificamos el abandono de las variables cualitativas para su aplicación. Es posible superar este problema con MDS, pues en vez de emplear una matriz de datos cuantitativos como input, se utiliza una **matriz de cuadrados de distancias**. Este método fue propuesto por Young & Householder (1938), redescubierto por Torgerson (1952) y Gower (1966).

El objetivo será obtener unos ejes o **Coordenadas Principales** $Y_1, \dots Y_m$ a partir de una matriz de distancias adecuada, donde sea posible realizar una representación euclídea que coincida con las distancias calculadas. Las principales dificultades asociadas con la aplicación de MDS son la interpretación de las Coordenadas Principales y la complejitud computacional.

Algorithm 1: Algoritmo de obtención representación MDS

Input: Matriz de datos mixtos $\mathbf{X}$ con variables cuantitativas $\mathbf{X}_1$ y cualitativas $\mathbf{X}_2$

Output: Coordenadas Principales $Y_1, Y_2, \dots, Y_m$

1. Ordenar la matriz de datos: $\mathbf{X} = [\mathbf{X}_1\, \mathbf{X}_2]$

2. Calcular:

$$\mathbf{M} = (\mathbf{x}_{i1} - \mathbf{x}_{j1})' \, \mathbf{S}^{-1} \, (\mathbf{x}_{i1} - \mathbf{x}_{j1}) \qquad \mathbf{C} = \alpha \oslash p$$

$$\mathbf{M}^{(2)} = \mathbf{M}^{\circ 1/2} \circ \mathbf{M}^{\circ 1/2} \qquad \mathbf{C}^{(2)} = 2(\mathbf{1}\mathbf{1}' - \mathbf{C})$$

$$\mathbf{M}^* = \mathbf{M}^{(2)} / \texttt{vgeom}(\mathbf{M}^{(2)}) \qquad \mathbf{C}^* = \mathbf{C}^{(2)} / \texttt{vgeom}(\mathbf{C}^{(2)})$$

$$\mathbf{D}^{(2)} = \mathbf{M}^* + \mathbf{C}^*$$

3. Construir:

$$\textbf{i)}\ \mathbf{G} = -\frac{1}{2}\mathbf{H}\mathbf{D}^{(2)}\mathbf{H} \ \ (\text{si } \mathbf{c}\mathbf{G} \geq 0 \ \ \forall \mathbf{c} \neq \mathbf{0})$$

$$\textbf{ii)}\ \tilde{\delta}_{ij}^2 = \begin{cases} \delta_{ij}^2 + c & \text{si} \quad i \neq j \\ 0 & \text{si} \quad i = j \end{cases} \quad \text{y volver a } \textbf{i)} \text{ (en caso contrario)}$$

4. Diagonalizar: $\mathbf{G} = \mathbf{U}\Lambda\mathbf{U}'$

5. Obtener: $\mathbf{Y} = \mathbf{U}\Lambda^{1/2}$

En el algoritmo previamente expuesto puede apreciarse cómo modificamos la Distancia de Gower al sumar matrices de cuadrados de distancias con una misma variabilidad geométrica. La primera de éstas, $\mathbf{M}^*$ mide la disparidad entre las variables cuantitativas, mientras que la segunda, $\mathbf{C}^*$, realiza lo mismo con las cualitativas. Nótese que hemos procedido a agrupar la variable binaria y multiestado de nuestro conjunto de datos dado que de otra forma se daba demasiado peso a éstas, llegando a perturbar considerablemente la visualización de la configuración MDS. Esto se debe a que la Distancia de Gower 'penaliza' las variables cuantitativas, siendo necesario para corregirlo un método de estimación robusta. En lo atinente a la distancia de las variables cuantitativas, se ha optado por el empleo de la Distancia de Mahalanobis pues es resiliente a los cambios de escala y no ignora la introducción de variables redundantes. Para las variables cualitativas comenzamos calculando la similaridad, dividiendo el número de atributos coincidentes entre observaciones con el total de variables explicativas (*matching coefficients*), y posteriormente se transforma a distancia estadística.

Hemos calculado asimismo la Distancia de Gower sin nuevas implementaciones al objeto de comparar ambas. Para ello es necesario imponer la condición de misma variabilidad geométrica, como propusieron Cuadras & Fortiana (1995), ya que si no cualquier ejercicio de paralelismo no resultaría del todo correcto.

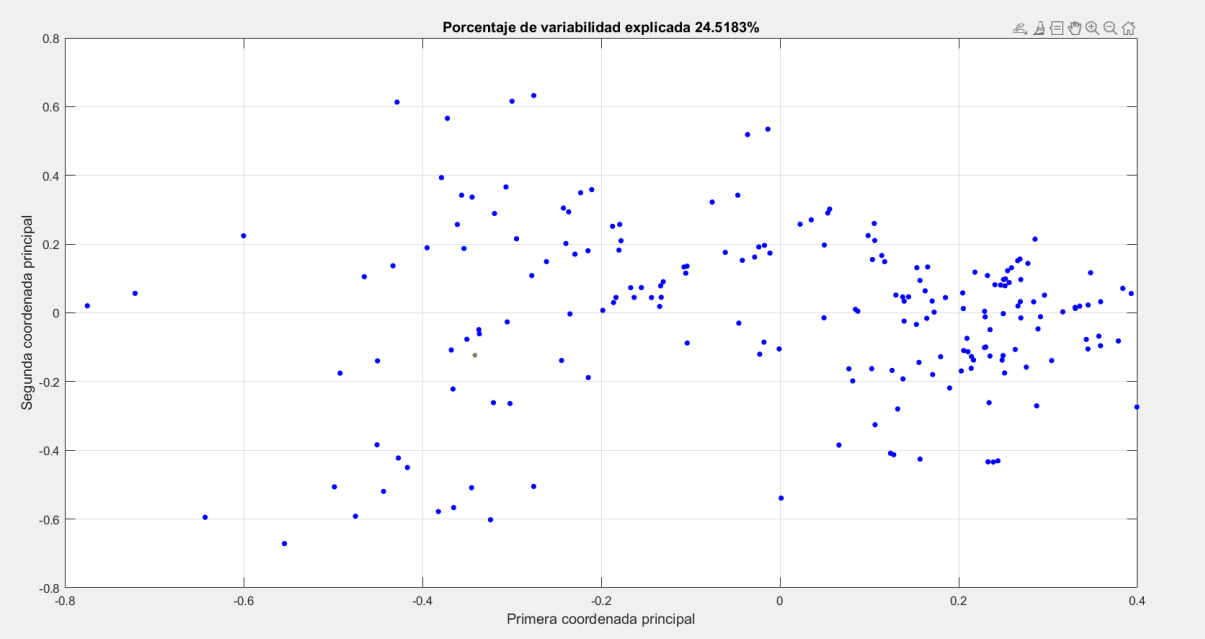

Figure 7: MDS a partir de Distancia de Gower propia (24.52 %)

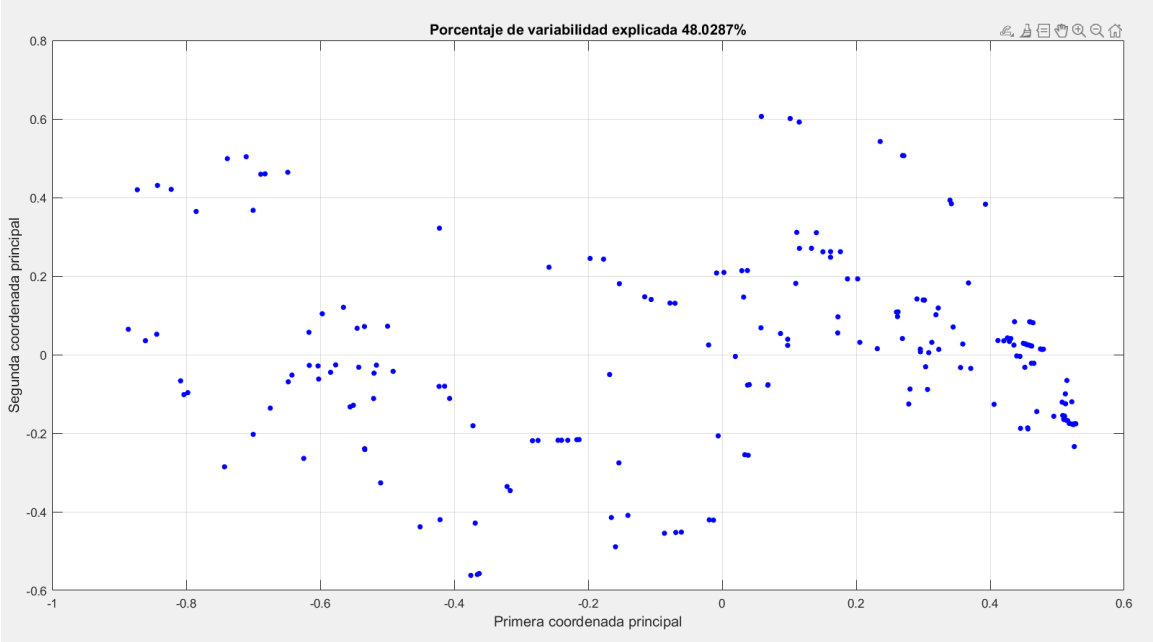

Figure 8: MDS a partir de Distancia de Gower (48.03 %)

La variabilidad explicada de la Distancia de Gower sigue una evolución no lineal, aumentando cada vez a menor ritmo. Aportamos gráfico de la Distancia que hemos propuesto. Se aprecia una tendencia de incremento de variabilidad explicada en una razón de 10 puntos porcentuales por cada autovalor, si bien decae a partir del noveno.

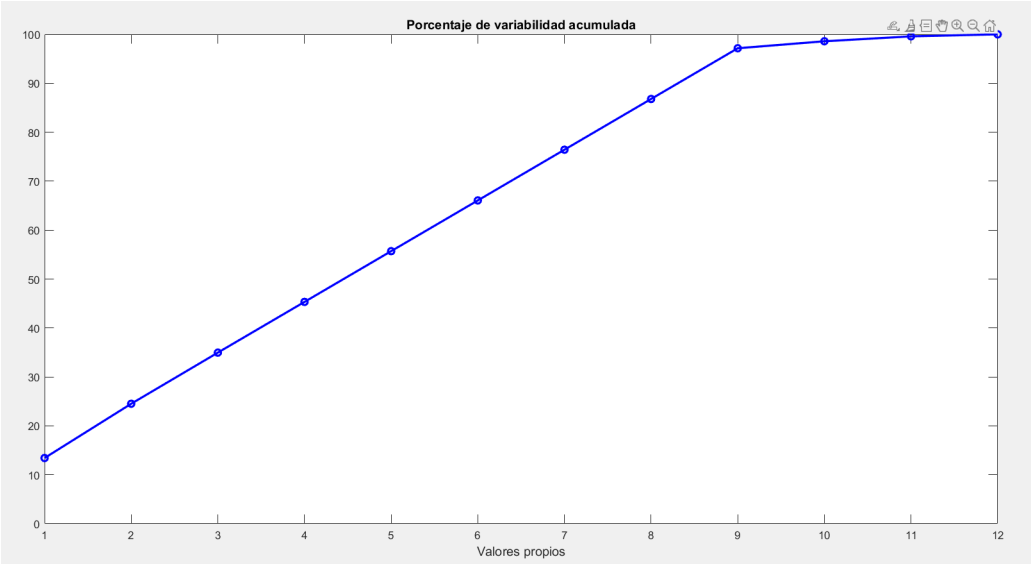

Figure 9: Porcentaje de variabilidad explicado

Empleando nuestra distancia observamos una mayor cantidad de puntos concentrados en los valores positivos del eje de abscisas (en adelante eje Y_1). La diferencia podría encontrarse en que la Distancia de Mahalanobis considera ciertos efectos que la Distancia Manhattan (utilizada por la Distancia de Gower) no podía superar.

Aunque en el caso de nuestra propia distancia la variabilidad explicada por las dos primeras Coordenadas Principales es considerablemente menor (24.52% contra un 48.03%), no deberíamos guiarnos por este criterio para decantarnos por una distancia u otra, pues se ignoran aspectos como una correcta definición de distancias, etc.

Resulta lógico que en un contexto de reducción de la dimensionalidad se descarte el criterio del 90% de variabilidad explicada como *ratio decidenci* del número de Coordenadas Principales (para nuestra distancia requeriríamos nueve), pues la práctica habitual es emplear dos o tres. En nuestro caso escogemos Y_1, Y_2, que explican respectivamente el 13.43% y 11.01% de la variabilidad de los datos. Excluimos el resto de Coordenadas Principales dado que la correlación que exhiben con los rasgos explicativos de nuestra matriz de datos es débil en prácticamente todos los casos (aportamos una prueba para Y_3 en la tabla).

Para la interpretación de las Coordenadas Principales nos serviremos de la relación entre éstas y las variables originales. Tenemos en cuenta a la hora de calcular correlaciones que existen diferentes tipos de variables. Así, emplearemos la correlación de Pearson para las variables cuantitativas, la V de Cramer para las binarias y la correlación de Spearman para el resto. La tabla con los valores correspondientes se muestra a continuación, así como la visualización en forma de mapa de calor.

Table 8: Correlaciones con Coordenadas Principales

Variables	Y_1	Y_2	Y_3
`distancia_ejes`	-0.6479	-0.2717	0.3626
`largo`	-0.6990	-0.1061	0.2298
`ancho`	-0.6521	-0.1696	0.3639
`altura`	-0.1269	-0.4378	-0.2436
`peso`	-0.8783	-0.0677	0.0569
`motor`	-0.6960	0.1272	0.2097
`caballos`	-0.7003	0.5172	-0.0019
`max_revoluciones`	0.2009	0.7209	-0.0396
`precio`	-0.8000	0.0980	0.1787
`combustible`	0.2350	0.5770	0.1500
`rueda_motriz`	0.7481	-0.1853	-0.0872

Nótese cómo la tercera Coordenada Principal apenas está correlacionada con las variables originales, de ahí que no se incluya en nuestro análisis para evitar cualquier interpretación desacertada o confusa. Procedemos con Y_1, Y_2 en adelante, constando que en lo sucesivo Y_2 denotará el eje de ordenadas para los comentarios sobre gráficos.

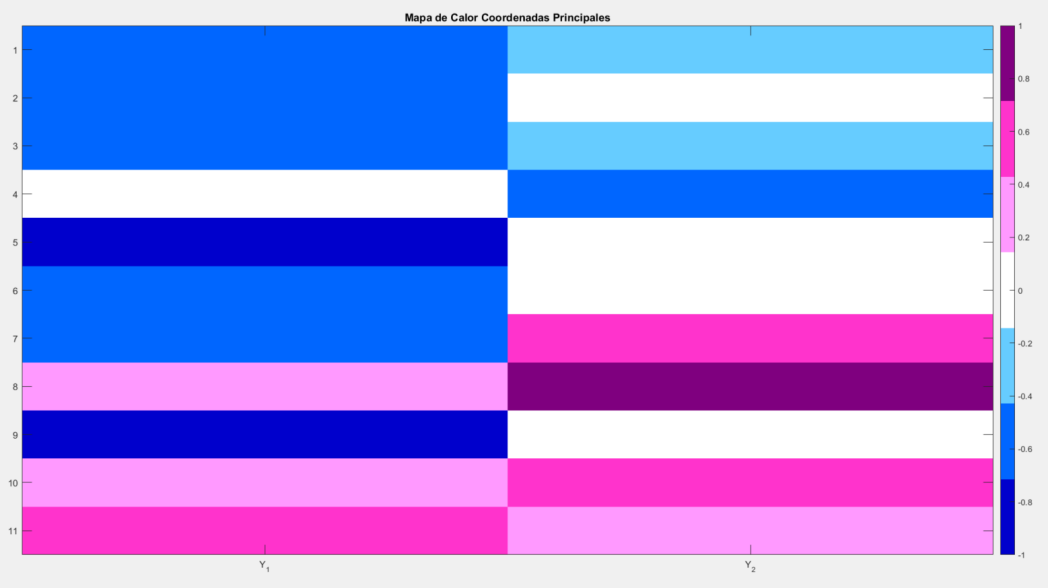

Figure 10: Correlación Coordenadas Principales y Variables

Las variables más correladas con Y_1 son (en orden): `peso`, `precio`, `rueda_motriz`, `caballos`, `largo`, `motor`, `ancho`, `distancia_ejes`, no jugando un papel relevante el resto (correlación por debajo de 0.3 en valor absoluto). El signo de la correlación es negativo en todos los casos salvo cuando se trata de la rueda motriz. En consecuencia, nos encontraremos a la izquierda del eje Y_1 del *gráfico de representación MDS* cuando el automóvil presente grandes dimensiones (coche pesado, ancho, con gran altura y distancia entre ejes), sea más caro y potente (muchos caballos de vapor y motor considerable). Podrán apreciarse más hacia la derecha los vehículos 4x4 y con tracción delantera. Por tanto es un índice de pequeñez y poca intensidad del coche, lo cual está asociado al precio.

En el caso de Y_2 las correlaciones son débiles menos para el caso de `max_revoluciones`. Le siguen `combustible`, `caballos` y `altura` (la única negativa de las mencionadas). En lo más alto del *gráfico de representación MDS* se situarán los coches de gran potencia (caballos y revoluciones altas) pero que no sobresalen en cuanto a la altura, premiando ligeramente los que usen gasolina. Podría entenderse como un índice de adecuación de los vehículos a la serie deportiva

Nótese que este tipo de coches presenta una gran potencia medida en caballos de vapor y revoluciones, así como una altura baja-media para incremental el aerodinamismo del vehículo (rasgos que identificamos en Y_2).

Ponemos en práctica la interpretación de las Coordenadas Principales que hemos sugerido de forma visual: presentamos una serie de gráficos donde se diferencian las variables de diversas formas (estudio de perfiles).

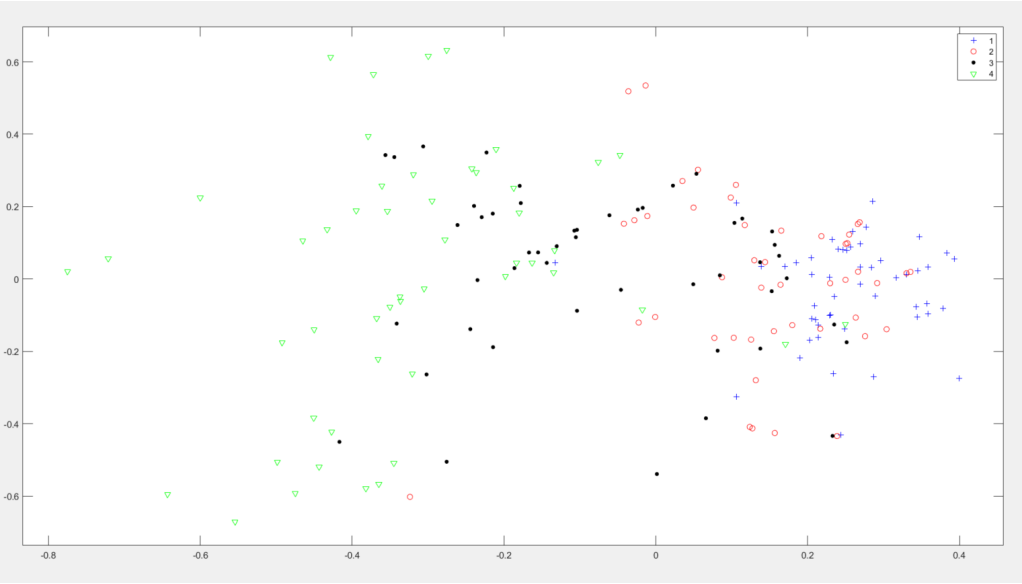

Figure 11: Precio por cuartiles

El gráfico de arriba agrupa los vehículos según su precio por cuartiles. Al objeto de que resulte más intuitivo, considérese el primer cuartil como modelos de gama baja de coches, el segundo cuartil como gama media-baja, tercer cuartil equivale a gama media-alta y último cuartil la gama alta. Recordemos a su vez que en el eje de abscisas, Y_1 o índice de pequeñez y poca intensidad, encontraremos automóviles de gran tamaño y vigor escorados a la izquierda, mientras que los coches más pequeños y de tecnología modesta estarán a la derecha. No resulta sorprendente que los coches de gama baja y media-baja (colores azul y rojo) presenten valores de Y_1 positivos. Asimismo, la gama alta (color verde) está significativamente presente en el extremo izquierdo del gráfico (valores negativos). Por último, la gama media-alta es más difícil de clasificar, mostrándose presente tanto en el lado negativo como positivo de los ejes (punto intermedio). En cuanto al eje Y_2, como ya comentamos, las observaciones con valores negativos están asociadas a vehículos distantes de encajar en el prototipo deportivo y que por ello son altos y poco potentes. De forma contraria, aquellos puntos

que observamos en el gráfico situados más hacia arriba, serán coches aerodinámicos y de gran intensidad. No es de extrañar que los automóviles con valores positivos considerables en el eje Y_2 asimismo se encuentran en el lado negativo de Y_1.

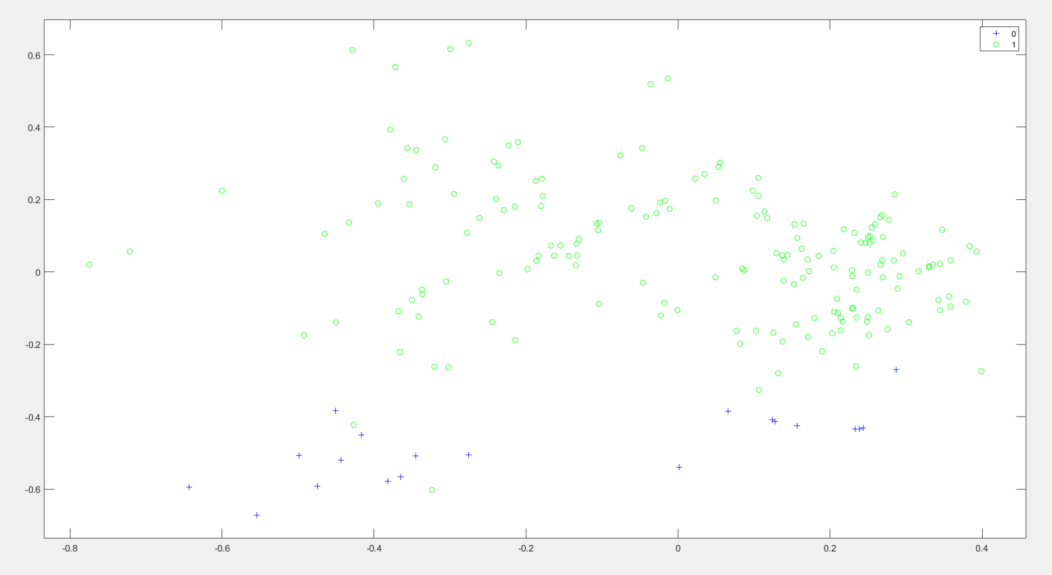

Figure 12: Representación MDS según combustible

Avanzando hacia el el análisis de las variables cualitativas, comenzamos por la parte binaria. En la Figura 8 se observa la representación MDS distinguiendo dos valores de `combustible`: gasolina (verde) y diésel (azul). La ligera mayoría de los automóviles diésel se caracterizan por su potencia y dimensiones, pues se localizan en la parte izquierda del eje Y_1, siendo de los valores más pequeños (indicativo de la capacidad del coche). Asimismo, los coches diésel se encuentran por debajo de los de gasolina en el eje Y_2, pues ya avanzamos que esta Coordenada Principal premiaba el uso de gasolina al presentar una asociación positiva. Cabe destacar que los automvviles con gasolina, al haber tantos y ser diversos, se encuentran dispersos a lo largo del eje Y_1, por lo que existen diferentes dimensiones y potencia. Es apreciable cómo el aerodinamismo es mayor en los coches de gasolina, consistente con el dato de que la mayoría de coches deportivos al tener tanta potencia, necesitan de una combustión directa y rápida, difícilmente compatible con la densidad del diésel (Auto10, 2018).

En cuanto a la variable multiestado, para los valores tracción trasera (coloz azul), tracción delantera (verde) y 4x4 (rojo) de `rueda_motriz`, observamos cómo los coches de tracción trasera son los más potentes y con mayor dimensión, seguidos de los 4x4 y por último los de las ruedas delanteras. Sobre el índice aerodinámico/deportivo o

Y_2 los todoterrenos no sobresalen como es esperable, los coches de tracción trasera son variados, así como los de tracción delantera. Señalamos como punto de mejora para éstos últimos la intensidad del motor y las revoluciones, existiendo espacio para mejorar en los aspectos ingenieriles y tecnológicos.

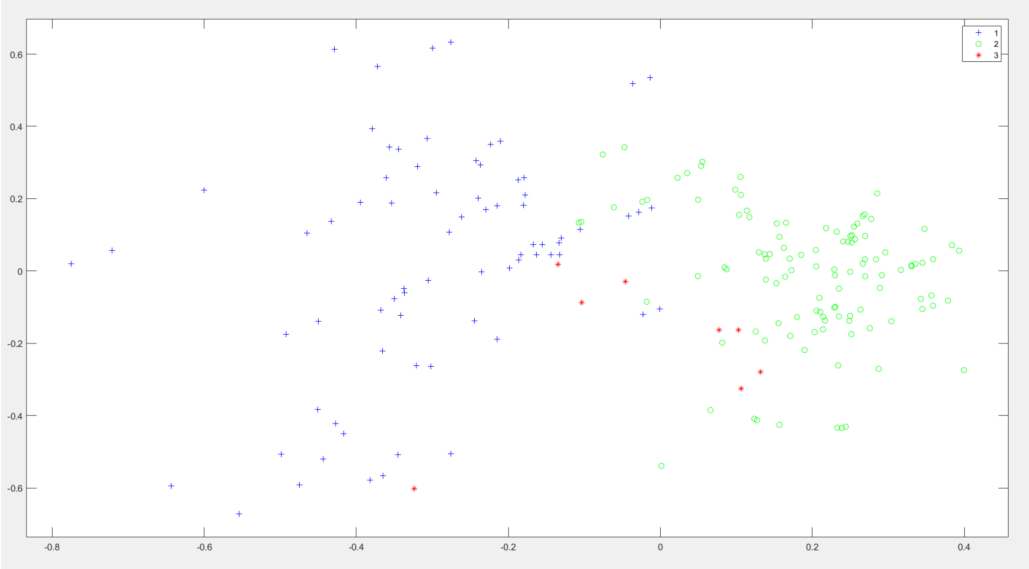

Figure 13: Representación MDS según tracción

Finalmente comentamos varios perfiles provenientes de dos variables cuantitativas, siendo la primera el peso del coche (Figura 9). Debido a que los cuatro grupos creados en el gráfico se identifican con los cuartiles, interprétese el color azul o primer cuartil como los coches ligeros, el colo rojo o segundo cuartil como automóviles de peso bajo-medio, el color negro o tercer cuartil como peso medio-alto y el color verde como vehículos pesados. El eje Y_1 presenta los coches pesados concentrados en el lado más negativo y los ligeros en el extremo derecho o valores más positivos. Resulta lógico que sea así pues Y_1 es un índice de pequeñez y el peso está positivamente asociado con la altura y ancho de los vehículos. En cuanto a los pesos intermedios, si bien es cierto que las observaciones atinentes al tercer cuartil están más proximas a los coches pesados, existe cierta dispersión que mezcla estos dos grupos. Concentrándonos en Y_2, resaltan los pesos medios como los más dinámicos, e incluso algunos coches pesados consiguen puntuar sorprendentemente bien en el índice de aerodinamismo (en parte se explica porque precisan de una mayor potencia para poder circular correctamente).

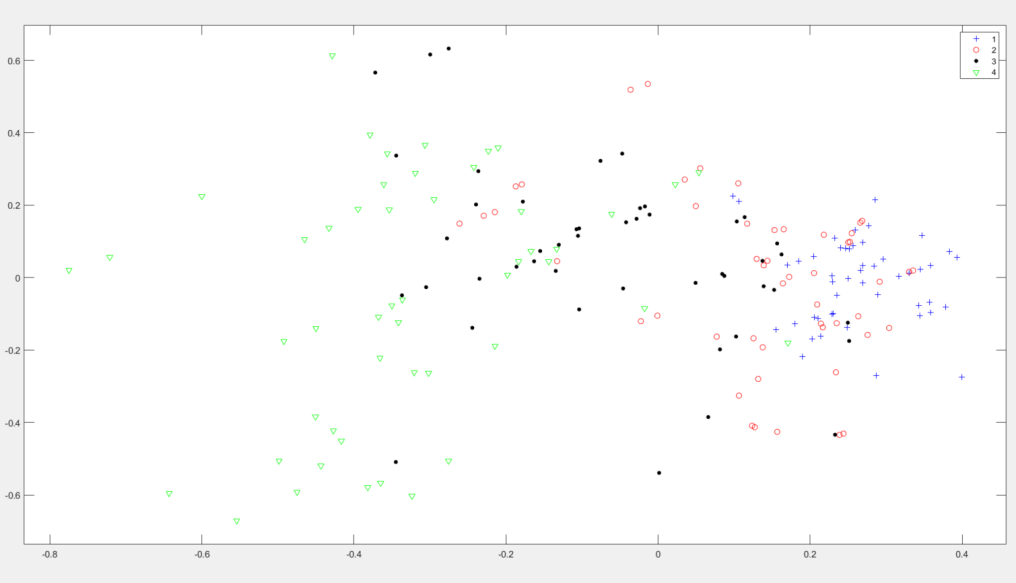

Figure 14: Representación MDS según peso

Concluyendo el proyecto con la variable motor: el color azul serían los motores más modestos, el rojo los mecanismos de automoción medios-pequeños, el color negro los motores medios-grandes y el color verde los de considerable dimensión. Observamos cómo los coches con motor grande lideran el índice de aerodinamismo (mayores valores en Y_2) ya que los deportivos cuentan con motores increíblemente potentes, requiriendo un mayor tamaño. Esto se traduce en muchos caballos de vapor y por ello encontramos estos mismos coches en los valores negativos del eje Y_1. Justamente lo contrario sucede con los automóviles con menor motor, y en cuanto a los que identificamos como intermedios experimentan una mezcla al encontrarse juntos.

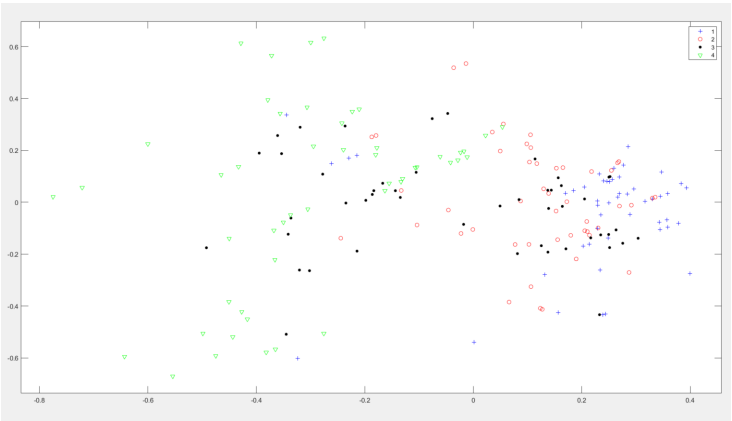

Figure 15: Representación MDS según motor

10 Clasificación Jerárquica

En esta sección aplicamos el método de Clasificación Jerárquica como primera técnica del Análisis de Conglomerados. Partimos de un escenario en el cual suponemos que en nuestro conjunto de datos no existe un etiquetado específico que determine el grupo al que pertenecen las observaciones. Consecuentemente, nos encontramos en un contexto de **clasificación no supervisada**, y por tanto emplearemos algoritmos que identifiquen y asignen automáticamente una observación dada a un *cluster* en el cual se encuentren otros individuos con características similares.

Obtendremos distintos conglomerados de manera sucesiva en clases de nivel superior que representaremos mediante un dendrograma. La base para lograr este objetivo es una matriz de distancias $\mathbf{D}^*$ que cumpla la propiedad **ultramétrica**:

$$\mathbf{D}^* \text{ es ultramétrica si } \begin{cases} \delta_{ij} = \delta_{ji} & \forall i, j \\ \delta_{ii} = 0 & \forall i \\ \delta_{ij} \leq \max\{\delta_{ik}, \delta_{kj}\} & \forall\, i, j, k \end{cases}$$

Emplearemos el siguiente algoritmo para obtener representaciones jerárquicas:

Algorithm 2: Algoritmo de tipo aglomerativo

Input: Matriz de distancias $\mathbf{D}^{(2)}$

Output: Matriz de distancias ultramétrica $\mathbf{D}^*$

1. Inicializar partición:

$$\varepsilon = \{1\} + \{2\} + \ldots + \{n\}$$

2. Crear un nuevo conglomerado:

$$\{i\} \cup \ \{j\} = \{i, j\} \quad \text{donde } \delta_{ij} = \min\{\delta_{kl}\}$$

3. Definir distancia de 2) al resto de elementos:

$$\delta'_{k,(ij)} = f(\delta_{ik}, \delta_{jk}) \quad k \neq i, j$$

$$\varepsilon = \{1\} + \cdots + \{i, j\} + \cdots + \{n\}$$

while $\varepsilon \neq \{1, 2, \ldots, n\}$ **do**

Pasos 2) y 3)

end

Con dicho algoritmo aunaremos iterativamente los automóviles y se recalcularán las distancias de los conglomerados del segundo paso al resto procurando cumplir la propiedad ultramétrica hasta obtener un único *cluster*. Nótese cómo la matriz que sirve de input es la Distancia de Gower personalizada que ***propusimos*** anteriormente, ya que nuestra base de datos presenta variables cuantitativas y cualitativas. Asimismo, es necesario señalar que el recálculo de distancias del algoritmo puede realizarse de diferentes modos, ya que mientras la función $f(\delta_{ik}, \delta_{jk})$ se defina de tal forma que no se comprometa la propiedad ultramétrica los resultados de los algoritmos de tipo aglomerativo son correctos. En concreto contrastaremos tres métodos diferentes de definir la función:

- Método del Mínimo: $f(\delta_{ik}, \delta_{jk}) = \min\{\delta_{ik}, \delta_{jk}\} \quad k \neq i, j$
- Método del Máximo: $f(\delta_{ik}, \delta_{jk}) = \max\{\delta_{ik}, \delta_{jk}\} \quad k \neq i, j$
- Método *Unweigthed Pair Group Method using Arithmetic Averages* (UPGMA): $\delta'(E_k, E_i \cup E_j) = \frac{n_i}{n_i+n_j}\delta E_i, E_k + \frac{n_J}{n_I+n_J}\delta E_j, E_k \quad k \neq i, j$

La representación gráfica de dichos métodos se aprecia a continuación:

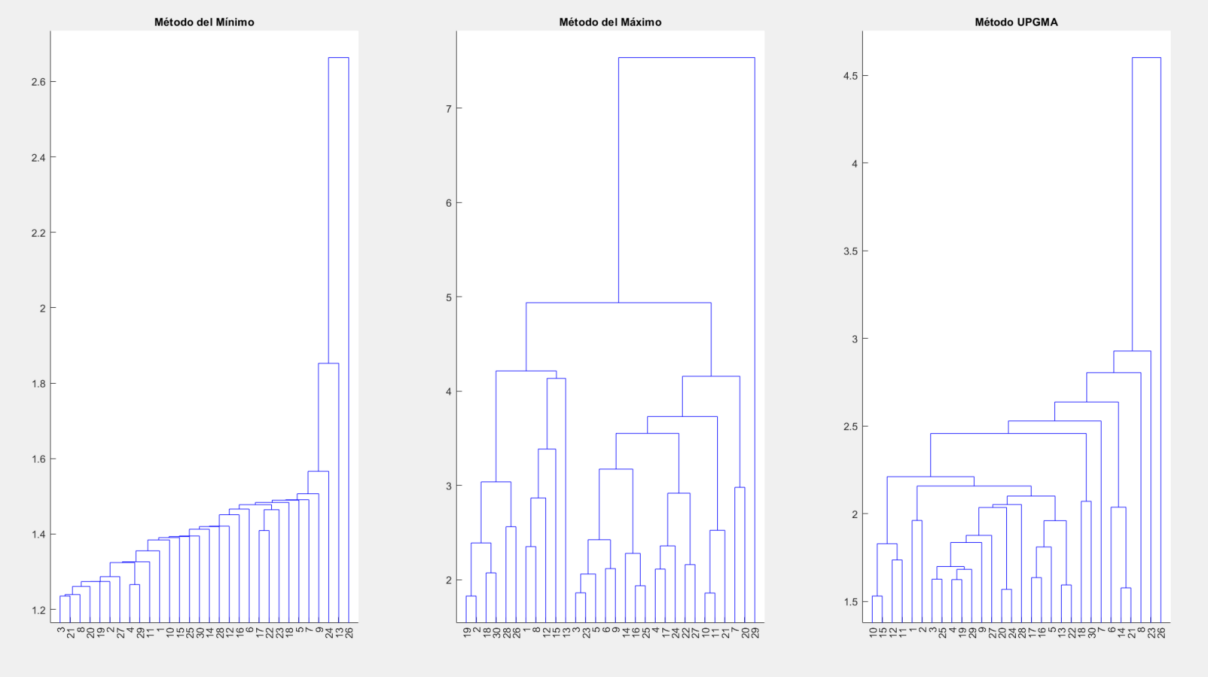

Figure 16: Comparación Clasificación Jerárquica

El primero de todos tiende a contraer los grupos, el segundo a dilatar el espacio de las clases y el último representa un punto más bien intermedio, dado que utiliza medias ponderadas según el número de elementos que hay en cada *cluster*.

Al objeto de elegir el método óptimo utilizaremos como métrica la Correlación Cofenética, un criterio de proximidad que determina el grado de alteración sufrido por la matriz de distancias original respecto a la nueva matriz ultramétrica. Mientras más cercana sea a uno menor es la alteración sufrida por $\mathbf{D}^{(2)}$ para convertirse en $\mathbf{D}^{*}$.

Una vez calculadas las Correlaciones Cofenéticas obtenemos que el Método UPGMA es el óptimo dado que es el que perturba menos nuestra matriz de distancias original (0.7899), seguida de cerca por el Método del Mínimo (0.7464), mientras que el Método del Máximo altera sustancialmente $\mathbf{D}^{(2)}$ con una Correlación Cofenética de 0.5166. A raíz de estos resultados es fácilmente perceptible que la matriz de distancias que hemos propuesto no cumplía la propiedad ultramétrica dado que ninguna Correlación Cofenética es igual a uno.

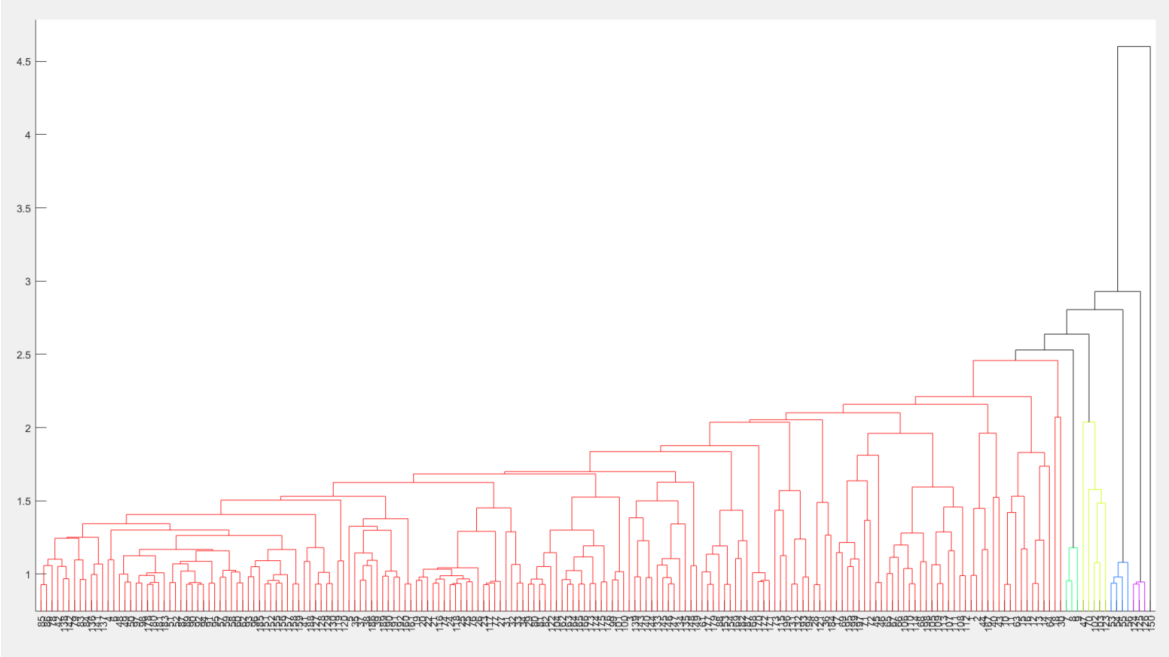

Figure 17: Método UPGMA

La interpretación de la clasificación jerárquica puede convertirse en una tarea intricada incluso para el caso de expertos en la materia de estudio, en buena medida explicado por la ausencia de una etiqueta que indique el grupo al que pertenecen las observaciones. No obstante, que no exista un agrupamiento previo y aún así se lo-

gre clasificar individuos es una evidente ventaja respecto de ténicas como la regresión logística o modelos probit entre otros, ya que precisan de un *label* o identificador de grupos para poder emplearse.

En un intento de explicar los resultados del Método UPGMA, en primer lugar cabe señalar cómo se interpreta el dendrograma: las ramas que se juntan antes (representadas por un mismo color) son más similares entre sí que el resto. Apreciamos cómo existe una distribución desigual entre *clusters* de la misma manera que nuestra base de datos presenta un pronunciado desequilibrio: 90% de los automóviles usan gasolina frente al 10% restante de diésel. Similarmente, el 58.7% de los coches utilizan tracción delantera, un 37% tracción trasera y sólo un 3.7% son todoterreno. Recordamos que los gráficos de configuración MDS que ***representamos*** en el anterior proyecto mostraban grupos claramente dominantes sobre otros (tendencias parecidas de distintos métodos).

Así, echando un vistazo detallado al dendrograma, los automóviles que pertenecen al conglomerado de color morado son de tracción trasera, su precio es el más alto de todos los grupos (en promedio 10.45 contra una media de 9.35) y también presentan los mayores valores para las variables `caballos` y `motor`. Es presumible que los coches más modernos, potentes y de estilo deportivo pertenezcan a este *cluster* según Zuto (2020). Por su parte los coches del grupo amarillo consumen gasolina, son de tracción trasera y su precio es mayor que la media, pero inferior a los anteriores. Dado que el peso de éstos es superior a los del grupo morado, es lógico concluir que aún situándose en una gama de calidad similar a éstos, los coches del conglomerado amarillo son modernos pero no deportivos.

El grupo mayoritario (rojo) incluye una verdadera amalgama de vehículos: todos los diésel y los todoterreno, si bien un análisis de medidas de centralidad revela que sus características se sitúan más bien en la media. Pese a representar el automóvil medio, acapara categorías que intuitivamente habrían de estar separadas en un principio. No debería de sorprendernos este resultado, ya que en tareas de clasificación es muy frecuente que grupos infrarrepresentados (como el caso de los diésel y todoterreno) no sean asignados con mucha precisión a su grupo correspondiente.

Y es que pese a haber realizado varias operaciones conducentes a aumentar la simetría de las variables, aproximar las unidades de medida y empleado una distancia de Gower un tanto más robusta, ni el preprocesamiento de los datos ni nuestro clasificador están cerca de ser calificados de sofisticados. Aún así hemos conseguido no perturbar considerablemente la matriz de distancias original e interpretar algunos *clusters*.

11 Clasificación No Jerárquica

En esta última sección aplicamos el algoritmo de k-means o k-medias, a veces referido como algoritmo Lloyd-Forgy, ya que fue propuesto por Stuart Lloyd en 1957 para tareas de modulación por impulsos codificados (no lo publicó fuera de la compañía Bell Laboratories hasta 1982), y Forgy (1965) por su parte compartió en *Biometrics* un procedimiento prácticamente idéntico.

Una explicación intuitiva del funcionamiento del algoritmo es la siguiente: en nuestro conjunto de datos elegimos k candidatos a representar un conglomerado (centroides), y asignamos las restantes observaciones a los *clusters* según estén más cerca de éstos. La complejidad computacional es generalmente lineal en relación al número de observaciones, el número de *clusters* y el número de dimensiones. La principal dificultad radica en que usualmente ni sabemos el número de centroides a elegir ni a cuál pertenece cada individuo de nuestra base de datos.

Por ello, es frecuente inicializar aleatoriamente los centroides, identificar las observaciones más cercanas a los mismos, actualizar los centroides en la siguiente iteración y proseguir sucesivamente hasta que los centroides no se muevan. La convergencia está asegurada ya que la distancia al cuadrado media de un determinado individuo y su centroide más cercano no puede más que disminuir por cada iteración, lo que no implica que sí converja a un óptimo global (Géron, 2019).

La plataforma computacional **MATLAB** emplea un método más sofisticado, concretamente el k-means++ propuesto por Arthur & Vassilvitskii (2007) en el que se tiende a seleccionar centroides distantes unos de otros, evitando más eficazmente soluciones subóptimas. La siguiente figura recoge el pseudo-código de los conceptos presentados:

Algorithm 3: Algoritmo de k-medias

Input: Número de clusters K, matriz de datos $\mathbf{X}$, tolerancia $\varepsilon > 0$

Output: Conjunto de K clusters

1. Inicializar centroides aleatoriamente: $\{\mu_1, \mu_2, \ldots, \mu_k\}$

while $\left||\mathbf{W}|^{(i+j)} - |\mathbf{W}|^{(i)}\right| > \varepsilon$ para $j \geq 1$ **do**

2. Por cada observación:

$$c^{(i)} = \arg\min_j \|x^{(i)} - \mu_j\|^2$$

$$\mu_j = \frac{\sum_{i=1}^{n} \mathbb{1}\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{n} \mathbb{1}\{c^{(i)} = j\}}$$

end

Nótese cómo se minimiza la matriz de dispersión dentro de los grupos para que las observaciones de un mismo conglomerado se asemejen lo máximo posible, convergiendo el algoritmo cuando dicha matriz no disminuya sustancialmente. Asimismo, es imprescindible recalcar que la medida utilizada para definir la separación de los datos a los *clusters* es la Distancia Euclídea, con todo lo que ello conlleva: restricción a variables cuantitativas, asumir incorrelación entre *inputs*, sensibilidad a cambios de escala... Consecuentemente, proponemos el siguiente procedimiento para superar las limitaciones mencionadas: obtener unos ejes de representación ortogonales que permitan incluir variables binarias y multiestado. Esto se conseguirá mediante el **Multidimensional Scaling** o **MDS** a partir de nuestra propia matriz de distancias al cuadrado $\mathbf{D}^{(2)}$ que introducimos en la sección anterior.

Aplicamos la función `kmedias2()` de Baíllo & Grané (2007) sobre la configuración MDS de nuestros datos y obtenemos las siguientes salidas gráficas:

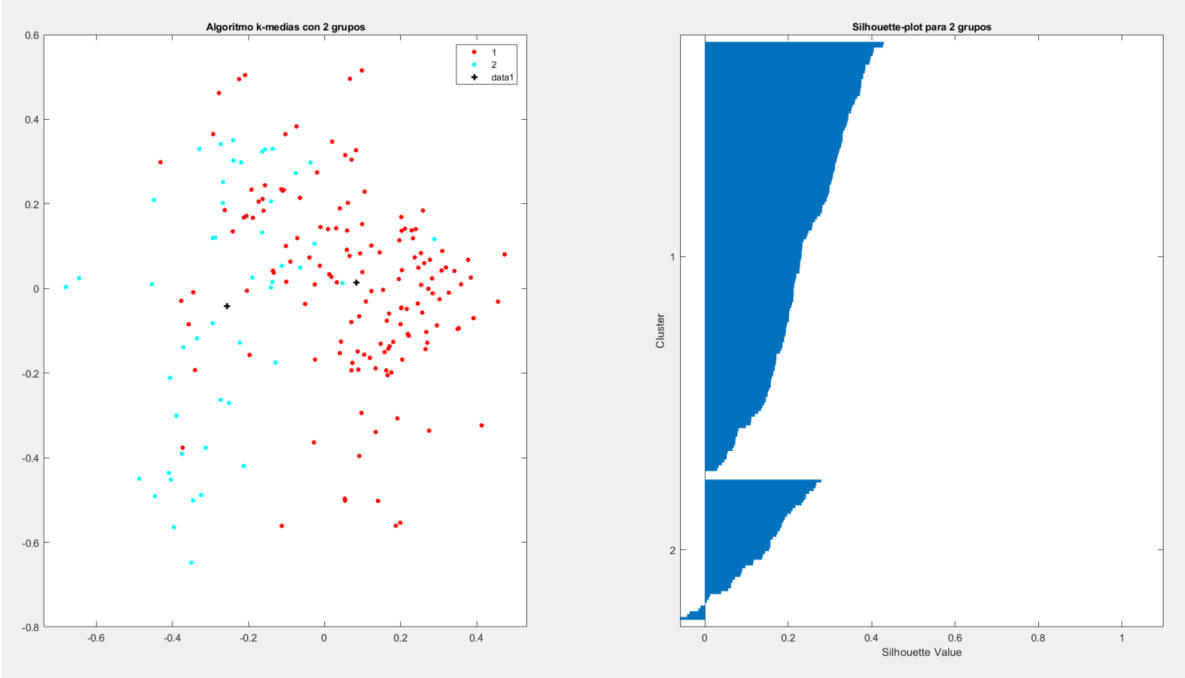

Figure 18: Algoritmo k-means (máximo 6 clusters)

El gráfico de la izquierda muestra la representación en dos dimensiones de la Clasificación del algoritmo k-means. A la derecha se aprecia el *Silhouette-plot* para evaluar la calidad del *clustering* a raíz de la siguiente métrica utilizada por Rousseeuw (1987):

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

La razón por la que no se ha empleado la técnica de PCA pese a procurar ejes incorrelados entre sí es fácilmente entendible a raíz de su propia naturaleza: sólo admite variables de tipo cuantitativo. Por ende, en caso de optar por PCA en vez de MDS estaríamos obviando la información contenida en los *inputs* de carácter cualitativo.

Volviendo a la métrica $s(i)$, considérese $a(i)$ como la distancia media de i al conglomerado que se asignó a dicha observación, y $b(i)$ como el mínimo de la distancia media de i al resto de *clusters*. Calculando la silueta media y considerando que en base a estas definiciones, $s(i) \in [-1, 1]$, el rendimiento de nuestro clasificador será alto si $\bar{s} \sim 1$ y todo lo contrario si se aproxima a -1. En nuestro caso la silueta media es 0.2134, lo que implica un resultado bastante mediocre, habida cuenta de que mientras más se acerque a cero la silueta media más incertidumbre rodea a nuestro clasificador.

El algoritmo de k-means detecta dos grupos en nuestro conjunto de datos y encuentra gran dificultad para identificar los componentes del segundo grupo (las únicas siluetas negativas se encuentran en dicho *cluster*), existiendo a su vez una gran vacilación por parte del clasificador en el resto de casos ($s(i) < 0.5 \ \forall i$).

Para interpretar el gráfico izquierdo, hará falta explicar el significado de las coordenadas principales Y_1, Y_2 a raíz de sus correlaciones con las variables originales (nos fijaremos en las relaciones mayores en valor absoluto). Y_1 presenta relaciones negativas con todas las variables menos las categóricas, por lo que lo tomaremos como un proxy inverso del tamaño (específicamente penaliza los coches con gasolina y de tracción trasera o 4x4). Y_2 exhibe correlaciones fuertes positivas únicamente para `precio` y `max_revoluciones`, de ahí que nos sirva como índice de potencia de los vehículos.

De esta manera, el primer grupo (color rojo), que se encuentra mayoritariamente en el lado derecho del eje de abscisas (valores positivos), agrupa a los automóviles de menor dimensión, con potencia más bien indefinida debido a la gran dispersión existente en el eje de ordenadas, si bien el centroide se identifica en la parte positiva de Y_2. Por ello, es esperable que coches de modesto tamaño pero con características técnicas heterogéneas se clasifiquen en el *cluster* rojo (encontraríamos deportivos y coches antiguos pequeños).

En cuanto al segundo conglomerado (color azul), lo identificamos en la parte negativa tanto de Y_1 como de Y_2 (sobre todo del primer eje), de ahí que esperemos coches más grandes pero que no optimicen tanto la tecnología de las revoluciones y sean además caros.

Aportamos una tabla con medidas de centralidad por grupos en la siguiente página.

Table 9: Grupos Algoritmo k-means

	Cluster Rojo			Cluster Azul		
Variables	Media	Mediana	Moda	Media	Mediana	Moda
`dist_ejes`	4.5653	4.5633	-	4.6719	4.6812	-
`largo`	5.1319	5.1399	-	5.2355	5.2407	-
`ancho`	3.3862	3.3869	-	3.4292	3.4348	-
`altura`	7.2805	7.2835	-	7.4868	7.4967	-
`peso`	7.7508	7.7456	-	8.0584	8.0488	-
`motor`	4.7153	4.6821	-	5.0568	5.0239	-
`caballos`	4.5011	4.4543	-	4.8200	4.8122	-
`max_revo`	8.5459	8.5564	-	8.5058	8.5172	-
`precio`	9.1977	9.0992	-	9.8240	9.7351	-
`combust`	-	-	1	-	-	1
`rueda`	-	-	2	-	-	1

Vemos cómo el *cluster* rojo o grupo 1 presenta valores medios y medianos menores que el *cluster* azul o grupo 2 en todas las variables cuantitativas a excepción de `max_revoluciones`, ya que como adelantamos los coches del segundo conglomerado no las optimizan tanto. Los coches con dimensión más reducida (valores inferiores de altura, anchura, largo, peso y distancia entre ejes), así como menor precio y caballos se encuentran en el primer grupo.

En cuanto a las variables cualitativas, en ambos grupos dominan los vehículos con gasolina (no sorprendente debido al gran desequilibrio de los datos) pero existen tendencias diferentes en cuanto a la tracción del coche: los automóviles más pequeños tienden a concentrar su fuerza en las ruedas traseras, mientras que los más grandes y caros lo hacen en las delanteras.

12 Conclusión

Las variables de nuestra base de datos muestran mayoritariamente una relación lineal positiva y fuerte. Las transformaciones no lineales nos han permitido acercarnos a una ley normal multivariante, supuesto necesario para los contrastes de hipótesis planteados. Ambos nos han permitido concluir que la evidencia disponible se muestra en contra de la igualdad del vector de medias entre los diferentes grupos estudiados.

Mediante la aplicación de técnicas de reducción de la dimensionalidad hemos sido capaces de crear diferentes grupos o perfiles que resumen la información contenida en

nuestra matriz de datos. Mientras PCA sólo es aplicable a las variables cuantitativas, MDS puede extenderse a todas. Concretamente hemos identificado que los índices más útiles han sido los de tamaño, modernidad y optimización de recursos (PCA) así como las dimensiones y fuerza del motor y aerodinamismo de los vehículos (MDS).

El principal objetivo ha sido mostrar el potencial de dichos métodos estadísticos, haciendo constar que su utilidad resaltaría aún más en contextos de bases de datos complejos. No obstante, ha de tenerse en cuenta que la interpretación de los resultados de PCA y MDS no es sencilla, recomendando en todo caso consultar con expertos cualquier duda o dificultad que se pueda encontrar para superar dicho obstáculo.

Hemos empleado dos métodos para realizar *clustering* no supervisado: clasificación jerárquica y no jerárquica. El primero crea conglomerados sucesivamente en clases de nivel superior mientras que el segundo parte de centroides y la distancia de las observaciones a éstos.

La calidad de los clasificadores difiere bastante: el Método UPGMA no ha perturbado demasiado la matriz original de distancias, mientras que la métrica de rendimiento del algoritmo k-means (silueta media) no ha sido abrumadora.

No obstante, hemos sido capaces de identificar tendencias y algunos patrones para el caso de la clasificación jerárquica e interpretar la conformación de grupos en el algoritmo k-means. Pese a que la Correlación Cofenética es alta en el caso de la Clasificación Jerárquica, intentar comprender sus resultados ha resultado ligeramente más difícil y no tan claro como el segundo caso.

Como limitación encontramos el gran desequilibrio existente en nuestra base de datos, así como ausencia de una posible mejor inicialización de los centroides en el caso de k-means. Al primer caso podría responderse con la aplicación de los algoritmos de remuestreo **SMOTE**, al segundo con recomendaciones de expertos que permitan cambiar de perspectiva: introducción de a prioris informativos para superar las limitaciones de la perspectiva frecuentista.

13 Apéndice

En el siguiente repositorio de GitHub puede encontrarse el código de ***MATLAB***, ***Python*** y ***R***.

Las funciones no modificadas y que no son propias de MATLAB se atribuyen a Baíllo & Grané (2007).

14 Referencias

Arturh, D., & Vassilvitskii, S. (2007). k-means++: The Advantages of Careful Seeding. *Conference: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*

Auto 10 (2018). *Diez pros y contras de los Coche Gasolina.* ***Enlace al artículo***.

Baíllo, A. & Grané, A. (2007). *100 Problemas Resueltos de Estadística Multivariante.* Delta Publicaciones

Ballard, D.H. (1987). Modular Learning in Neural Networks. *Association for the Advancement of Artificial Intelligence, 1987 Conference.*

Cuadras, C. M. & Fortiana, J. (1995). A Continuous Metric Scaling Solution for a Random Variable. *Journal of Multivariate Analysis*

Forgy, E.W. (1965). Cluster Analysis of Multivariate Data: Efficiency vs Interpretability of Classifications. *Biometrics*

Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn and TensorFlow.* O'Reilly

Gower, J.C. (1966). Some Distance Properties of Latent Root and Vector Methods Used in Multivariate Analysis. *Biometrika*

LeasePlan (2017). *Coche de Combustión vs Coche Eléctrico: ¿Cuál Gana?* ***Enlace al artículo***.

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*

Motorpasion (2012). *El motor de combustión es el más eficiente hoy: FALSO* ***Enlace al artículo***.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*

Torgerson, W.S (1952). Multidimensional scaling: I. Theory and method. *Psychometrika*

Young, G. & Householder, A. S. (1938). Discussion of a set of points in terms of their mutual distances. *Psychometrika*

Zuto (2020). *The Car Size Evolution.* Zuto Car Finance