# 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** para llevar a cabo tareas de Data Cleaning, Análisis Exploratorio de Datos y Estadística Multivariante, al fin de preparar adecuadamente las variables que servirán de base para la obtención de resultados relevantes. Las operaciones siguen una lógica que iremos exponiendo a lo largo del proyecto.

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 **201** 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.

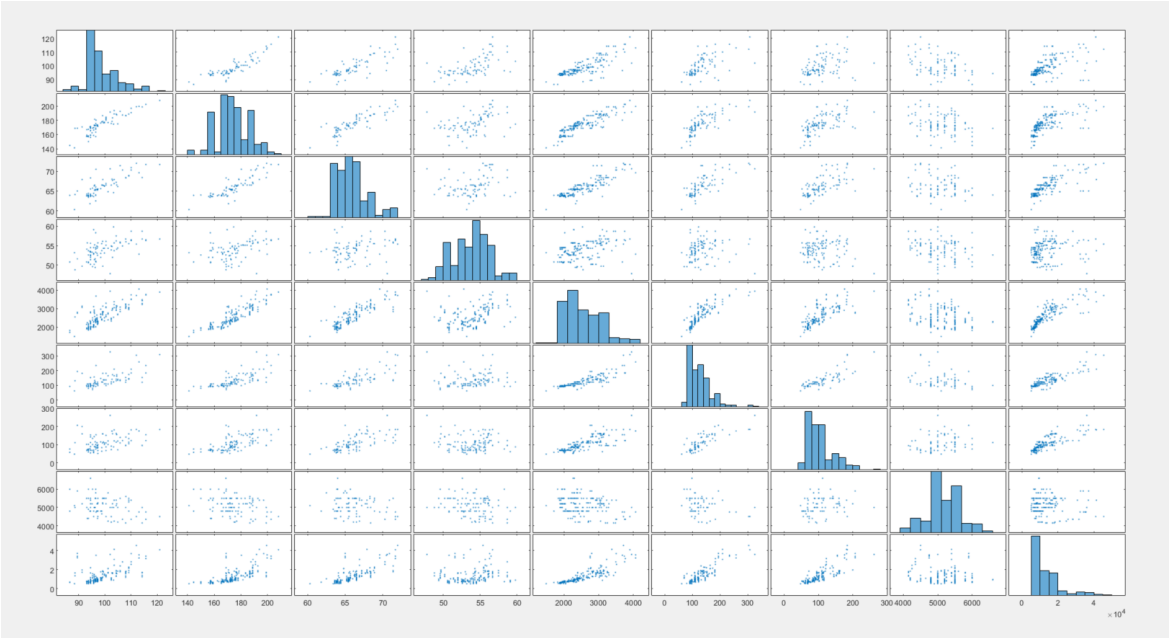

Figure 1: Gráfico de Dispersión Matricial Original

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.

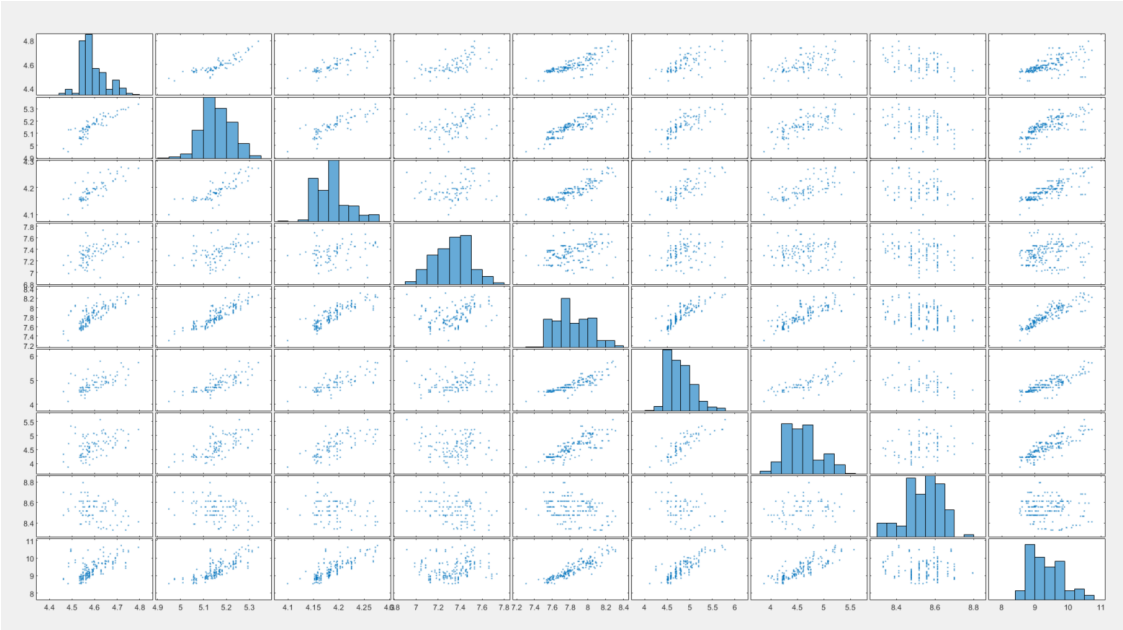

Figure 2: Gráfico de Dispersión Matricial con Variables Transformadas

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 $\eta^2$.

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 $\eta^2$ 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 |
| $\eta^2$ | 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:

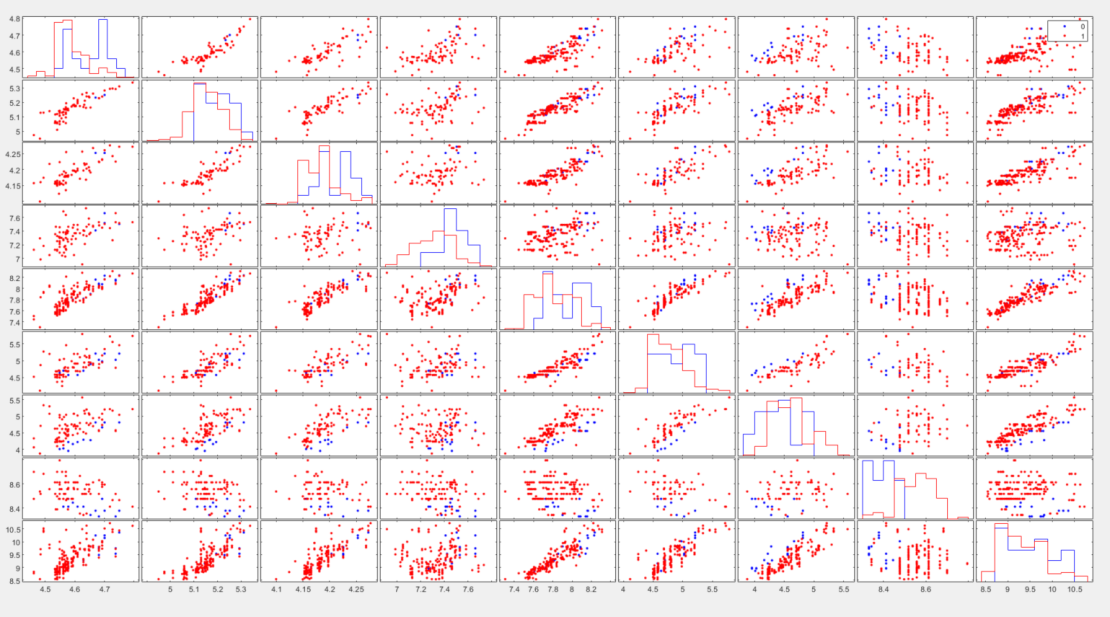

Figure 3: Gráfico de Dispersión Matricial con Variables Transformadas

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 tiene filas provenientes 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`.

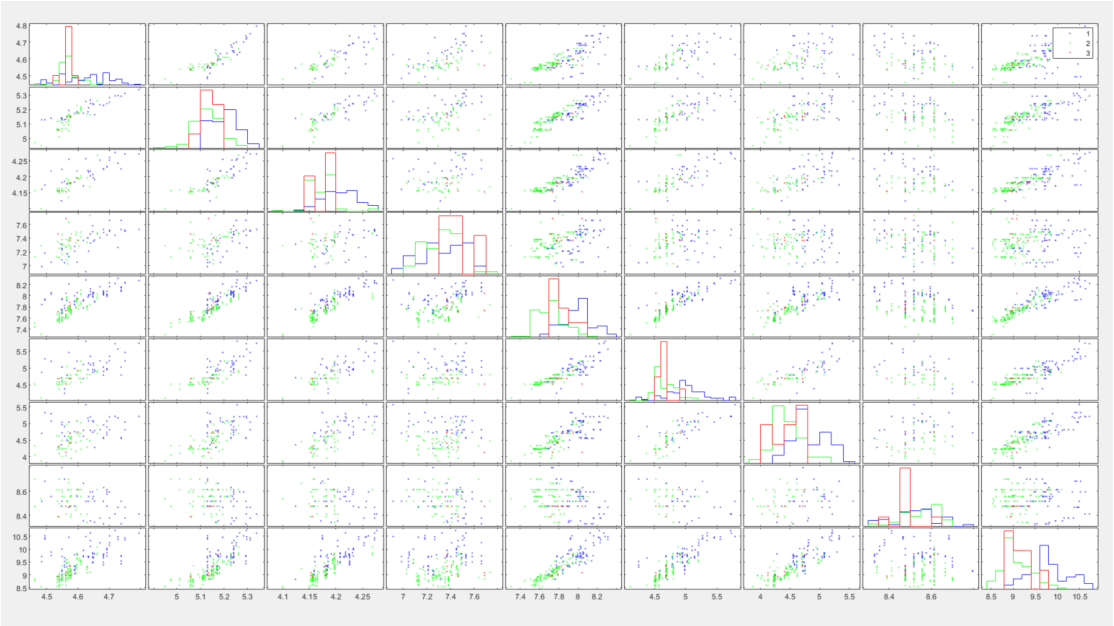

Figure 4: Gráfico de Dispersión Matricial con Variables Transformadas

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 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.

# 8 Apéndice

El código de Python utilizado puede encontrarse en el siguiente ***Notebook***.
El código de MATLAB y R puede encontrarse en el siguiente repositorio de ***GitHub***.