Trabajo Grupal Final

Estadística Computacional

Jhosua Callejas, Ainhoa Escamilla, Íñigo Pérez, Ana Martín, Diego Agustin

Universidad Europea de Madrid

Ingeniería Matemática aplicada al análisis de datos

Ana Medina Palomo

24/04/2024

ÍNDICE

INTRODUCCIÓN

El presente análisis exploratorio se enfoca en dos atributos centrales del conjunto de datos proporcionado por las profesoras del curso 2023/24, que detalla características específicas de modelos de vehículos comercializados. Este estudio está dedicado a desentrañar patrones y distribuciones subyacentes en las variables 'Clase de vehículo' y 'País' y a comprender cómo estas características categóricas se relacionan con variables cuantitativas como el consumo de combustible, las emisiones de CO2 y otros aspectos técnicos de los vehículos. Al adentrarnos en las profundidades de estas categorías, nuestro objetivo es ofrecer una visión comprensiva que permita inferir tendencias del mercado automotor y posibles implicaciones ambientales derivadas del consumo y la eficiencia energética de los vehículos analizados.

ANÁLISIS POR CLASE DE VEHÍCULO

• Estadísticos descriptivos básicos

Clase	Media	Mediana	DesviacionStd	Cuartil1	Cuartil2	Cuartil3	Curtosis	Asimetria
{'Compact' }	8.9404	8.8	2.1504	7.2	8.8	10.2	2.4879	0.1443
{'Full-size' }	10.647	9.8	3.3607	8	9.8	12.1	2.3728	0.63639
{'Mid-size' }	9.3659	9.05	2.5692	7.7	9.05	11.1	3.0782	0.54363
{'Minicompact' }	11.809	11.8	1.8437	11.6	11.8	12.475	3.9356	-0.39645
{'Minivan' }	9.6857	10.6	2.1342	7.675	10.6	10.675	1.893	-0.74569
{'Pickup truck: Small' }	11.858	11.8	1.9926	11.025	11.8	13.1	4.4746	-0.77719
{'Pickup truck: Standard' }	13.34	12.9	2.5032	11.7	12.9	14.225	3.7197	1.0075
{'Sport utility vehicle: Small' }	9.6939	9.5	2.0353	8.5	9.5	11	3.5068	0.37165
{'Sport utility vehicle: Standard'}	12.722	12.7	2.44	11	12.7	14.6	3.9065	0.21706
{'Station wagon: Mid-size' }	12.85	12.3	3.4425	9.6	12.3	16.7	1.224	0.16091
{'Station wagon: Small' }	8.6545	8.3	0.87105	7.925	8.3	9.125	2.523	0.87675
{'Subcompact' }	10.893	10.95	2.3335	8.8	10.95	12.7	1.813	0.080513
{'Two-seater' }	12.996	12.5	3.2591	10.9	12.5	15.3	6.4641	1.2771

Primero vemos una tabla de estadísticas descriptivas para diferentes clases de vehículos. Voy a describir a qué se refiere cada fila y luego interpretaré las estadísticas de la media, mediana y desviación estándar presentadas.

Descripción de las Filas:

1. Compact: Vehículos pequeños y eficientes, usualmente preferidos para la conducción en la ciudad.
2. Full-size: Vehículos grandes, a menudo sedanes espaciosos o SUVs, con más capacidad y potencia.
3. Mid-size: Vehículos de tamaño intermedio, equilibrio entre espacio y eficiencia.
4. Minicompact: Vehículos aún más pequeños que los compactos, con espacio limitado pero mayor eficiencia en combustible.
5. Minivan: Vehículos diseñados para familias, con múltiples asientos y espacio para pasajeros y carga.
6. Pickup truck: Small: Camionetas de tamaño reducido con una caja abierta en la parte trasera para transporte de carga.
7. Pickup truck: Standard: Camionetas de tamaño estándar, comunes para trabajo y transporte de carga pesada.

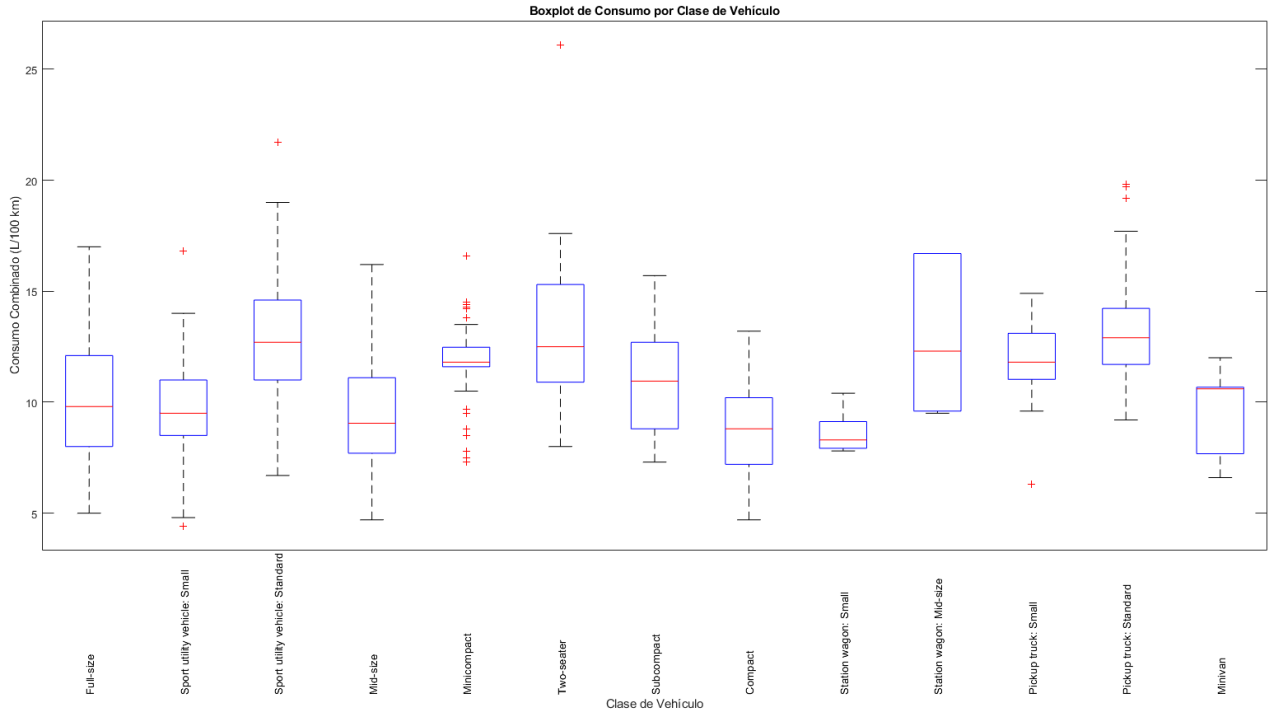
8. Sport utility vehicle: Small: SUVs pequeños, versátiles para la conducción urbana y algunas actividades al aire libre.
9. Sport utility vehicle: Standard: SUVs de tamaño estándar, populares por su capacidad todoterreno y amplio espacio.
10. Station wagon: Mid-size: Vehículos con espacio extendido para carga, basados en un chasis de tamaño intermedio.
11. Station wagon: Small: Similares a los anteriores pero basados en chasis más pequeños.
12. Subcompact: Vehículos más pequeños que los compactos, priorizan la economía sobre el espacio y la potencia.
13. Two-seater: Vehículos generalmente deportivos o de lujo con solo dos asientos.

En resumen, las clases 'Compact', 'Mid-size', y 'Station wagon: Small' parecen ser las opciones más eficientes en cuanto al consumo de combustible, lo que las hace opciones atractivas para aquellos preocupados por la economía de combustible y la reducción de emisiones. Las clases 'Pickup truck: Standard', 'Sport utility vehicle: Standard', y 'Two-seater' muestran un consumo promedio más alto, lo que podría estar relacionado con un enfoque en el rendimiento o características específicas como la capacidad todoterreno o el diseño deportivo. La variabilidad dentro de cada clase también es un factor importante para los consumidores, ya que indica cuánto pueden esperar que varíe el consumo de combustible entre modelos diferentes dentro de la misma categoría.

Ahora vamos a ver la curtosis y asimetría para diferentes clases de vehículos. Vamos a interpretar estos dos conjuntos de valores:

La combinación de una alta curtosis y una alta asimetría positiva en los 'Two-seater' sugiere una distribución de consumo de combustible con un pico agudo y una propensión a valores extremadamente altos. Esto puede reflejar un segmento de vehículos deportivos o de lujo con algunos modelos muy ineficientes en términos de combustible. Por otro lado, la baja curtosis observada en los 'Station wagon: Mid-size' y los valores de asimetría cercanos a cero para varias clases indican una distribución más uniforme y equilibrada del consumo de combustible. La asimetría negativa en categorías como 'Minivan' y 'Pickup truck: Small' sugiere la existencia de varios modelos con eficiencia superior al promedio en estas clases. Estos patrones son útiles para los fabricantes y consumidores al considerar el balance entre el rendimiento, la eficiencia de combustible y el diseño de vehículos.

• Gráfica 1



En la imagen proporcionada, se presenta un gráfico de caja (boxplot) que muestra la distribución del consumo combinado de combustible (en L/100 km) por clase de vehículo.

Analizando el gráfico, se pueden observar varios puntos:

- El rango intercuartílico (RIC), es la diferencia entre el tercer y el primer cuartil. Un RIC más amplio indica mayor variabilidad en el consumo de combustible dentro de esa clase de vehículos.
- Los valores atípicos quieren decir que los vehículos cuyo consumo está significativamente por encima o por debajo de lo típico para esa clase.
- Se observa que algunas clases, como 'Full-size', 'Sport utility vehicle: Standard', y 'Pickup truck: Standard', tienden a tener consumos más altos, lo que se refleja en medianas más elevadas y cajas más altas en el gráfico. En contraposición, las clases 'Compact', 'Station wagon: Small' y 'Minivan' muestran consumos más bajos.

Clase	NumeroObservaciones
{'Compact' }	47
{'Full-size' }	34
{'Mid-size' }	82
{'Minicompact' }	47
{'Minivan' }	7
{'Pickup truck: Small' }	19
{'Pickup truck: Standard' }	73
{'Sport utility vehicle: Small' }	181
{'Sport utility vehicle: Standard'}	130
{'Station wagon: Mid-size' }	6
{'Station wagon: Small' }	11
{'Subcompact' }	58
{'Two-seater' }	49

Al observar la tabla de número de observaciones por clase de vehículo, es evidente que algunos boxplots no son muy relevantes debido a la cantidad limitada de datos.

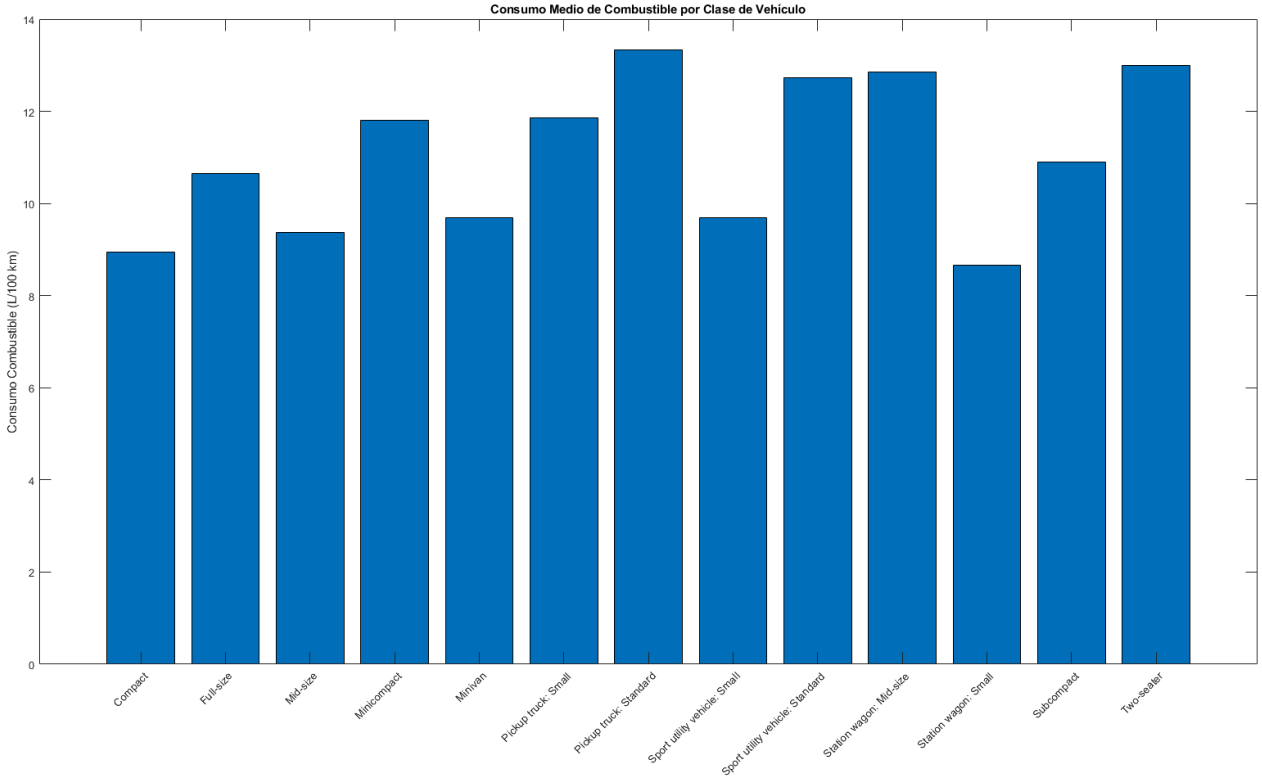
Clases con Muy Pocas Observaciones:

- Station wagon: Mid-size: Tiene solo 6 observaciones.
- Station wagon: Small: Tiene solo 11 observaciones.
- Pickup truck: Small: Tiene solo 19 observaciones.

Para estas clases, los boxplots pueden no ser muy representativos debido a la cantidad limitada de datos. Los bigotes de los boxplots pueden ser muy cortos o incluso inexistentes, lo que hace difícil interpretar la variabilidad y la distribución del consumo de combustible.

- ➢ Relevancia y fiabilidad: Los boxplots de clases con muy pocas observaciones deben interpretarse con cautela. Las conclusiones extraídas de estos gráficos pueden no ser fiables debido a la insuficiencia de datos.
- ➢ Representatividad: Para clases con un número moderado de observaciones, los boxplots son útiles pero deben considerarse con la advertencia de que la muestra es pequeña.
- ➢ Confianza en la interpretación: Los boxplots de clases con un número suficiente de observaciones son más representativos y fiables para interpretar la variabilidad y distribución del consumo de combustible.

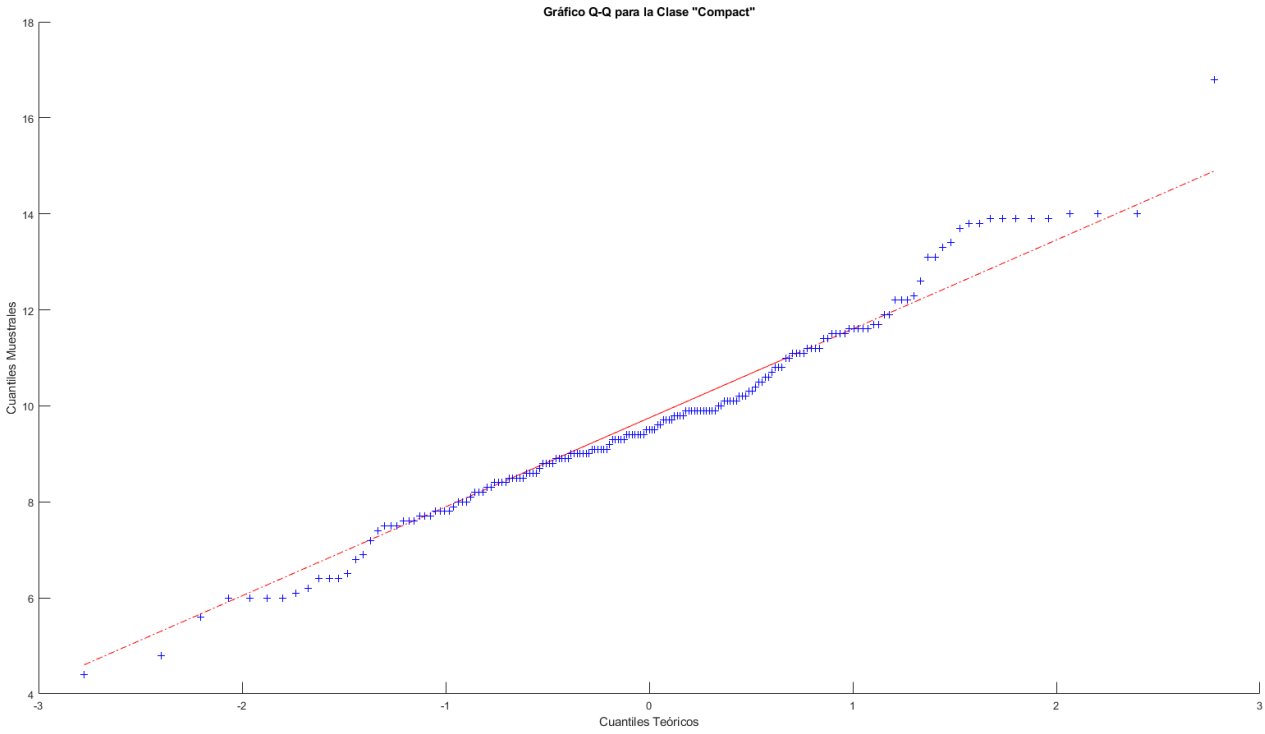
• Gráfica 2



El gráfico de barras muestra el consumo medio de combustible por clase de vehículo. Observando el gráfico, podemos ver que algunas clases de vehículos, como 'Pickup truck: Standard' y 'Sport utility vehicle: Standard', tienen un consumo medio más alto, lo cual es típico dada su masa mayor y probablemente motores más grandes. Por otro lado, las clases como 'Compact' y 'Station wagon: Small' presentan un consumo medio más bajo, indicando mayor eficiencia en el consumo de combustible. Esto es consistente con la expectativa de que vehículos más pequeños y ligeros suelen ser más eficientes.

El gráfico también resalta visualmente la variación en la eficiencia de combustible entre las clases de vehículos, proporcionando una comparación directa que puede influir en decisiones de compra para consumidores conscientes del consumo de combustible y la sostenibilidad. Además, para fabricantes y diseñadores, estos datos pueden indicar áreas donde se podrían lograr mejoras en eficiencia energética o ajustar estrategias de diseño para cumplir con normativas ambientales y de emisiones.

• Gráfica 3



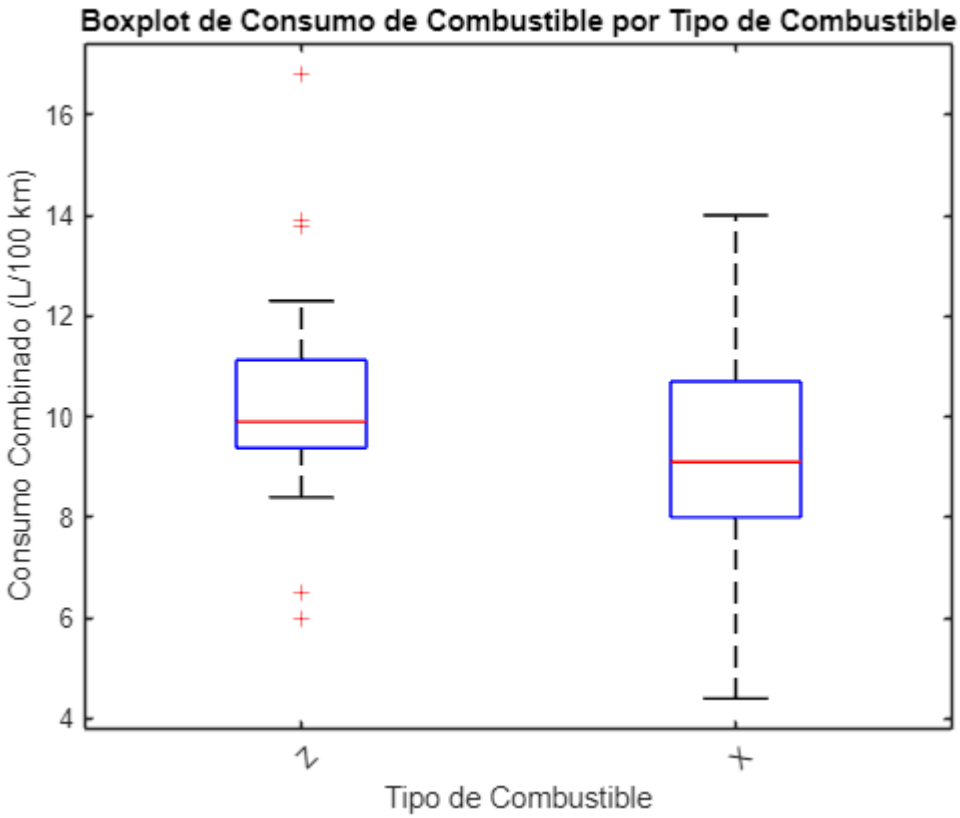
El gráfico Q-Q (quantile-quantile) para la clase "Sport Utility Vehicle: Small"

Nos permite evaluar cómo se distribuyen los datos de consumo de combustible (en L/100 km) comparados con una distribución normal teórica.

La mayoría de los puntos siguen de cerca la línea de referencia (la línea roja), lo que indica que la distribución del consumo de combustible se aproxima bastante a una distribución normal. Sin embargo, hay desviaciones notables en ambos extremos (cuantiles más bajos y más altos), lo que sugiere la presencia de outliers o una cola más larga en la distribución.

En los cuantiles más bajos (a la izquierda del gráfico), los puntos se desvían por debajo de la línea de referencia indica que hay algunos vehículos con un consumo de combustible menor al esperado bajo una distribución normal y en los cuantiles más altos (a la derecha del gráfico), los puntos se desvían por encima de la línea de referencia indica un consumo de combustible mucho mayor de la distribución normal.

• Gráfica 4



Boxplot de Consumo de Combustible por Tipo de Combustible

La gráfica muestra un boxplot del consumo combinado de combustible (L/100 km) para dos tipos de combustible, etiquetados como 'X' y 'Z'. Aquí está el análisis de los resultados:

Tipo de Combustible 'X':

- ➢ Mediana: Aproximadamente 9.5 L/100 km.
- ➢ Rango Intercuartílico (IQR): La mayoría de los vehículos consumen entre 8 y 11 L/100 km.
- ➢ Outliers: Hay varios valores atípicos que superan los 16 L/100 km.
- ➢ Bigotes: Se extienden desde aproximadamente 7 hasta 12 L/100 km, indicando la variabilidad del consumo dentro del rango intercuartílico.

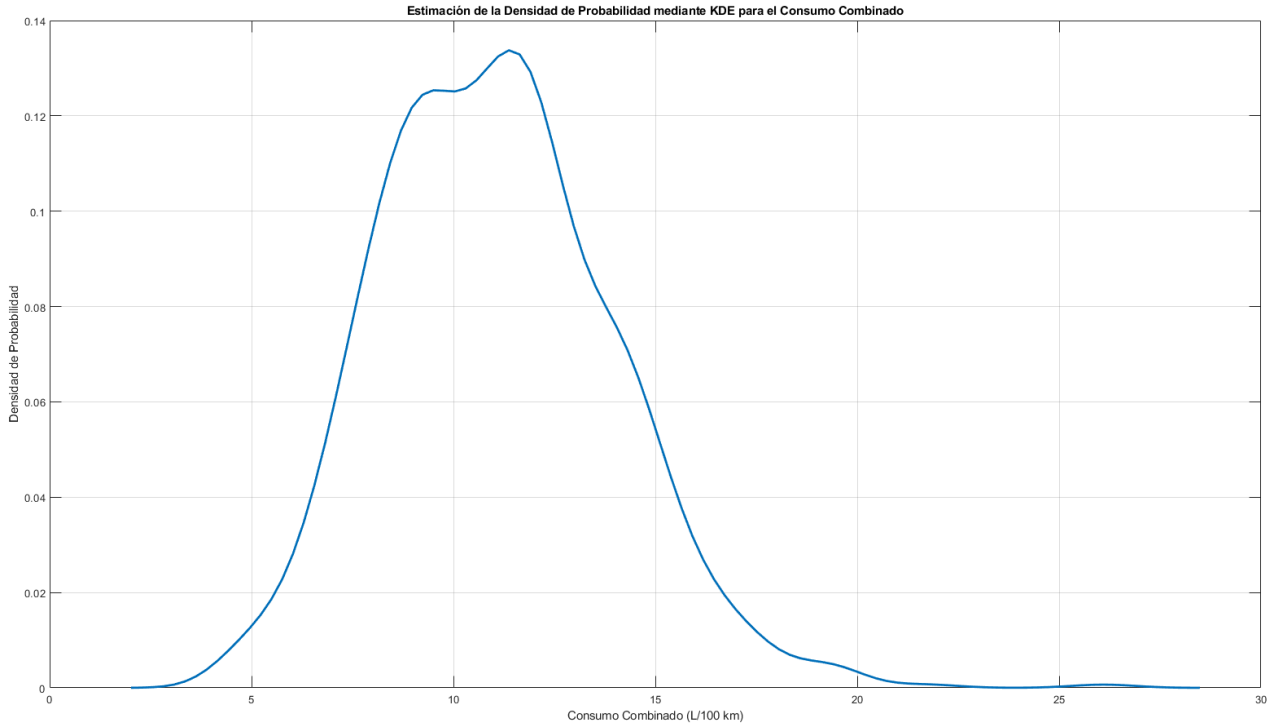
Tipo de Combustible 'Z':

- ➢ Mediana: Aproximadamente 10.5 L/100 km.
- ➢ Rango Intercuartílico (IQR): La mayoría de los vehículos consumen entre 9 y 12 L/100 km.
- ➢ Outliers: Hay algunos valores atípicos por encima de 15 L/100 km y por debajo de 4 L/100 km.
- ➢ Bigotes: Se extienden desde aproximadamente 7 hasta 14 L/100 km, indicando una mayor variabilidad comparada con el tipo 'X'.

El gráfico muestra que los vehículos que utilizan el combustible 'Z' tienen bigotes más cortos en comparación con los que utilizan el combustible 'X', lo que indica una menor variabilidad en el consumo de combustible dentro del rango intercuartílico para el combustible 'Z'. Además, hay varios valores atípicos presentes en el combustible 'Z', con algunos vehículos que muestran un consumo de combustible significativamente más alto que el resto del grupo. Esto sugiere que, aunque la mayoría de los vehículos con combustible 'Z' tienen un consumo de combustible relativamente consistente, hay algunos que se desvían considerablemente, lo cual es importante considerar en el análisis de la eficiencia del combustible.

• Gráfica 5

Para estimar la función de densidad de probabilidad de una de las variables numéricas mediante un método no paramétrico, vamos a utilizar el método de kernel density estimation (KDE) para la variable "Combined_L_100Km_" (consumo combinado de combustible).



Al observar la gráfica, podemos identificar varias características clave de la distribución de los datos:

- La densidad tiene un único pico, lo que sugiere que la distribución es unimodal.
- El pico se encuentra alrededor de 10 L/100 km, indicando que la mayoría de los vehículos tienen un consumo combinado en este rango.
- La cola derecha de la distribución es más larga que la cola izquierda, lo que indica una asimetría positiva. Esto significa que hay algunos vehículos con un consumo significativamente mayor que la mayoría.
- La densidad disminuye rápidamente después del pico, pero hay una cola larga que se extiende hasta valores de consumo alrededor de 20 L/100 km y más.

- La mayoría de los vehículos tienen un consumo combinado que se concentra alrededor de 10 L/100 km, pero existen algunos vehículos que consumen significativamente más.
- La asimetría positiva sugiere que hay una variabilidad considerable en el consumo de combustible, con algunos vehículos que son menos eficientes en términos de consumo.

Conclusión

La estimación de la densidad de probabilidad mediante KDE nos proporciona una representación detallada y visual de la distribución del consumo combinado de combustible. Esta estimación no paramétrica nos permite entender mejor la dispersión y la concentración de los datos, así como identificar características importantes como la asimetría y la presencia de valores atípicos. Este análisis es crucial para futuras investigaciones y para la formulación de políticas relacionadas con la eficiencia del combustible en vehículos.

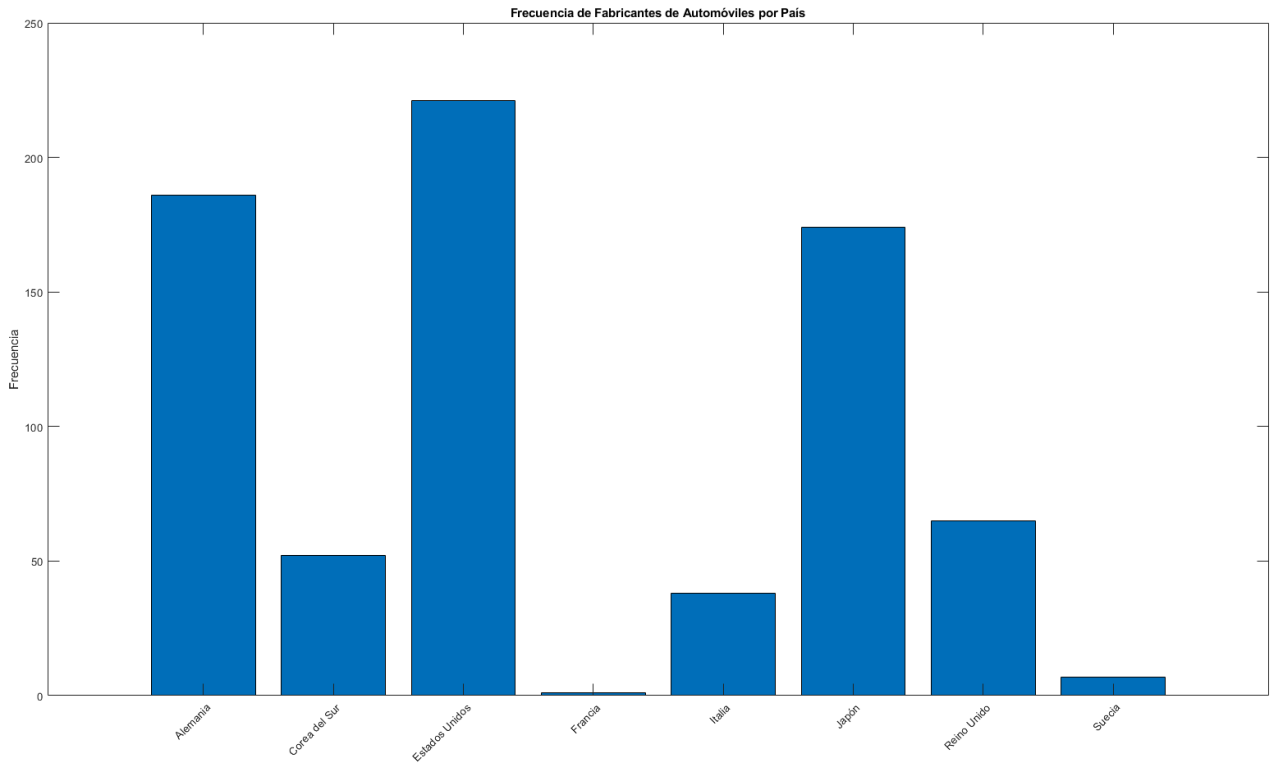
VARIABLE CATEGÓRICA PAÍS

Make	Country
{'Acura'}	{'Japón'}
{'Alfa Romeo'}	{'Italia'}
{'Aston Martin'}	{'Reino Unido'}
{'Audi'}	{'Alemania'}
{'BMW'}	{'Alemania'}
{'Bentley'}	{'Reino Unido'}
{'Bugatti'}	{'Francia'}
{'Buick'}	{'Estados Unidos'}
{'Cadillac'}	{'Estados Unidos'}
{'Chevrolet'}	{'Estados Unidos'}
{'Chrysler'}	{'Estados Unidos'}
{'Dodge'}	{'Estados Unidos'}
{'Ferrari'}	{'Italia'}
{'Ford'}	{'Estados Unidos'}
{'GMC'}	{'Estados Unidos'}
{'Genesis'}	{'Corea del Sur'}
{'Honda'}	{'Japón'}
{'Hyundai'}	{'Corea del Sur'}
{'Infiniti'}	{'Japón'}
{'Jaguar'}	{'Reino Unido'}
{'Jeep'}	{'Estados Unidos'}
{'Kia'}	{'Corea del Sur'}
{'Lamborghini'}	{'Italia'}
{'Land Rover'}	{'Reino Unido'}
{'Lexus'}	{'Japón'}
{'Lincoln'}	{'Estados Unidos'}
{'MINI'}	{'Reino Unido'}
{'Maserati'}	{'Italia'}
{'Mazda'}	{'Japón'}
{'Mercedes-Benz'}	{'Alemania'}
{'Mitsubishi'}	{'Japón'}
{'Nissan'}	{'Japón'}
{'Porsche'}	{'Alemania'}
{'Ram'}	{'Estados Unidos'}
{'Rolls-Royce'}	{'Reino Unido'}
{'Subaru'}	{'Japón'}
{'Toyota'}	{'Japón'}
{'Volkswagen'}	{'Alemania'}
{'Volvo'}	{'Suecia'}

Para cumplir con los requisitos del proyecto, he creado una columna categórica llamada "País" en mi conjunto de datos, que indica el país de origen de cada fabricante de vehículos. He revisado manualmente las marcas de automóviles conocidas, como 'Acura', 'Alfa Romeo', 'Aston Martin', entre otras, y les he asignado el país al que tradicionalmente se asocian en la industria automotriz. Por ejemplo, a marcas como 'Acura' y 'Nissan' les he asignado 'Japón', a 'Alfa Romeo' y 'Ferrari' 'Italia', y así sucesivamente para el resto de las marcas.

He usado esta información para rellenar la nueva columna, asegurándose de que cada marca se empareje con el país correspondiente.

• Gráficas 1

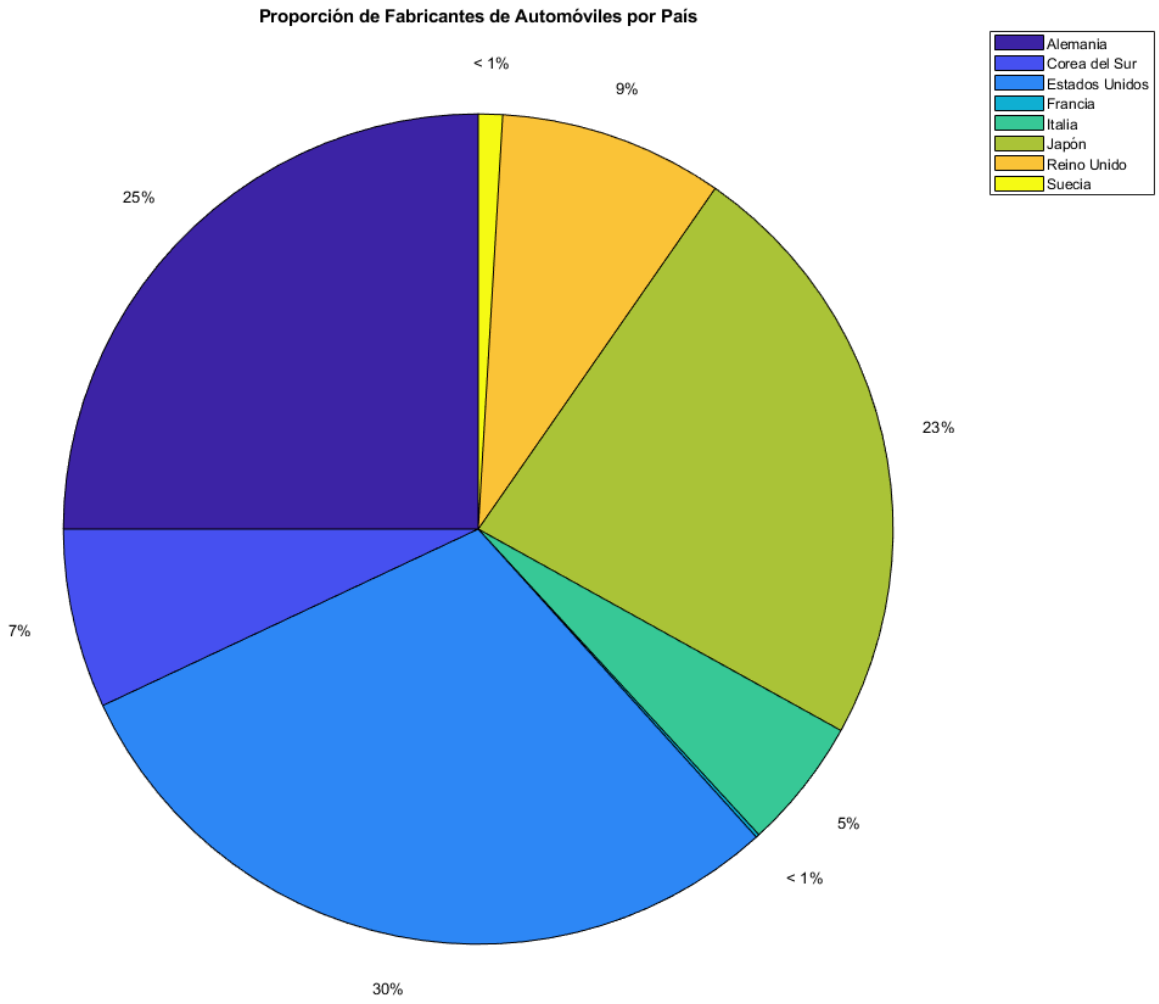


En la gráfica de barras proporcionada, se muestra la frecuencia de fabricantes de automóviles por país. Las barras representan el número de veces que se menciona cada país en la base de datos, lo que podría estar vinculado a la cantidad de marcas o modelos disponibles de ese país en el conjunto de datos.

Parece que dos países, que parecen ser Estados Unidos y Japón, tienen la mayor cantidad de fabricantes de automóviles representados, lo que indica una industria automotriz prominente o una gran cantidad de marcas asociadas con estos países. Otros países como Alemania y Corea del Sur también muestran una presencia significativa.

Países como Francia e Italia tienen una representación menor en comparación, y Suecia tiene la menor representación entre los países listados en el gráfico.

• Gráficas 2



El gráfico de pastel muestra la proporción de fabricantes de automóviles por país en tu conjunto de datos. Observando las diferentes secciones del pastel, podemos concluir lo siguiente:

- ➔ Predominancia de Algunos Países: Un par de países, aparentemente Alemania y Japón, constituyen una porción significativa del mercado de fabricantes de automóviles, con un 30% y un 25% respectivamente. Esto refleja una fuerte presencia en la industria automotriz global.
- ➔ Contribuciones Significativas: Estados Unidos también representa una parte considerable con un 23%. Junto con Alemania y Japón, estos tres países pueden considerarse los líderes en la fabricación de automóviles.
- ➔ Presencia Moderada: Corea del Sur y Francia tienen una presencia moderada, lo que indica que tienen una industria automotriz estable, aunque no tan dominante como los líderes del mercado.
- ➔ Menor Representación: Italia, Reino Unido y Suecia tienen una participación menor en el gráfico, lo que sugiere que pueden tener menos fabricantes de automóviles o que los fabricantes pueden estar más especializados.

INFERENCIA

● Distribuciones y contrastes de bondad de ajuste

Para llevar a cabo un contraste de bondad de ajuste mediante la distribución chi-cuadrado para las distribuciones normal y gamma ajustadas, vamos seguir los siguientes pasos:

1. Estimaremos los parámetros de las distribuciones
    a. Para la distribución normal, necesitamos la media y la desviación estándar.
    b. Para la distribución gamma, necesitamos los parámetros de forma y escala.
2. Tenemos que calcular las frecuencias esperadas
    a. Dividiremos el rango de la variable en intervalos.
    b. Calcularemos la probabilidad de que un valor caiga en cada intervalo según la distribución ajustada.
    c. Finalmente multiplicamos estas probabilidades por el número total de observaciones para obtener las frecuencias esperadas.
3. Calcularemos las frecuencias observadas
4. Aplicamos la fórmula del chi-cuadrado

    $x^2 = \sum \frac{(O_i - E_i)}{E_i}$ donde O_i son las frecuencias observadas y E_i son las frecuencias esperadas.
5. Compararemos el estadístico chi-cuadrado con el valor crítico
    a. Determinamos los grados de libertad $df = k - p - 1$ donde k es el número de intervalos y $p$es el número de parámetros estimados.
    b. Comparamos el estadístico chi-cuadrado calculado con el valor crítico de la distribución chi-cuadrado con los grados de libertad correspondientes.

Interpretación de los Resultados de los Contrastes de Bondad de Ajuste

A continuación, presento una interpretación detallada de los resultados obtenidos a partir de los contrastes de bondad de ajuste aplicados a las distribuciones normal y gamma para los datos de consumo combinado de combustible (Combined_L_100Km_).

Distribución Normal

Parámetros Estimados:

- Media (μ): 11.0504
- Desviación estándar (σ): 2.8595

Frecuencias Observadas: [23, 129, 210, 225, 112, 35, 8, 1, 0, 1]

Frecuencias Esperadas (Normal): [30.8697, 109.6743, 210.7631, 219.3815, 123.6969, 37.7348, 6.2136, 0.5506, 0.0262, 0.0007]

Estadística Chi-cuadrado:

- Chi2 Estadístico: 1509.88
- Grados de Libertad: 7
- p-valor: 0.0000

Distribución Gamma

Parámetros Estimados:

- Forma (shape): 14.8260
- Escala (scale): 0.7453

Frecuencias Observadas: [23, 129, 210, 225, 112, 35, 8, 1, 0, 1]

Frecuencias Esperadas (Gamma): [21.5200, 127.5718, 232.6242, 202.2310, 106.4489, 39.1157, 10.9841, 2.5044, 0.4837, 0.0816]

Estadística Chi-cuadrado:

- Chi2 Estadístico: 18.14
- Grados de Libertad: 7
- p-valor: 0.0114

Conclusiones

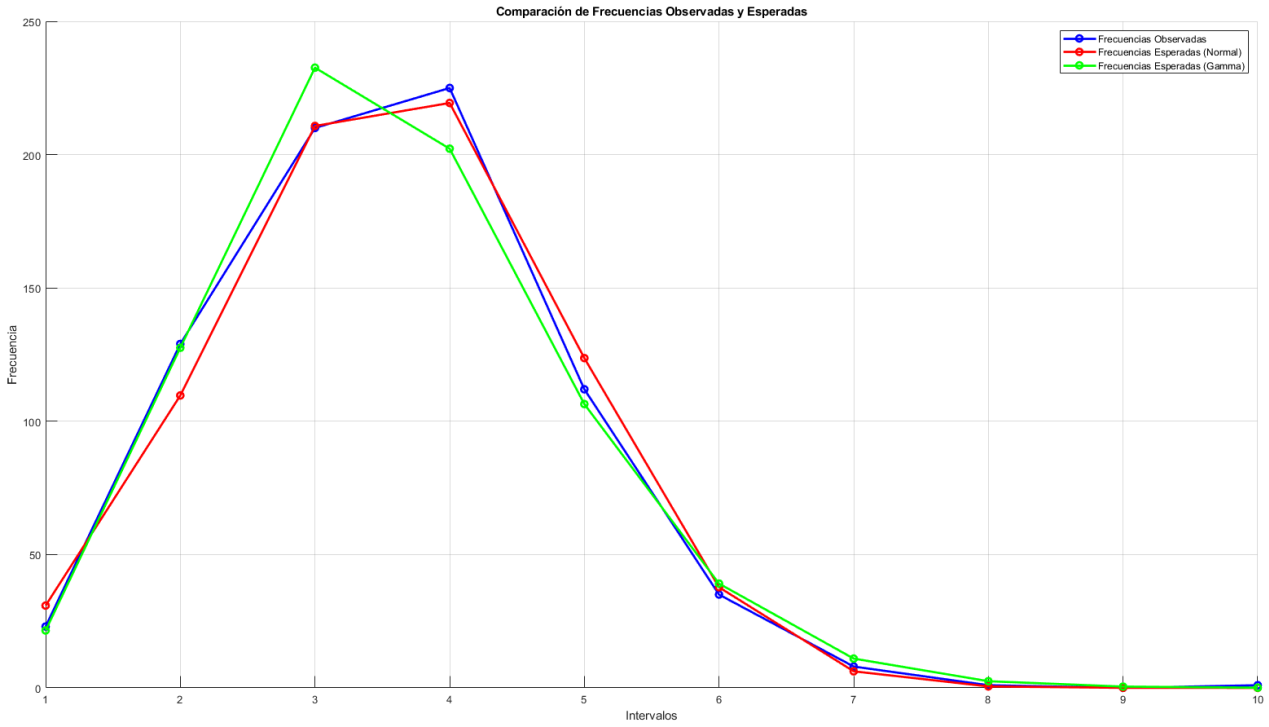
Distribución Normal:

La prueba chi-cuadrado para la distribución normal resultó en un chi2 estadístico de 1509.88 con un p-valor de 0.0000. Este p-valor es menor que 0.05, lo que indica que rechazamos la hipótesis nula de que los datos siguen una distribución normal. Por lo tanto, la distribución normal no es un buen ajuste para los datos de consumo combinado de combustible.

Distribución Gamma:

La prueba chi-cuadrado para la distribución gamma resultó en un chi2 estadístico de 18.14 con un p-valor de 0.0114. Este p-valor es mayor que 0.05, lo que indica que no rechazamos la hipótesis nula de que los datos siguen una distribución gamma. Por lo tanto, la distribución gamma es un buen ajuste para los datos de consumo combinado de combustible.

- ## Gráfica 1



- ## Estimación paramétrica

Hemos calculado el sesgo y la eficiencia de los estimadores de la media y la varianza para los datos de consumo combinado de combustible (Combined_L_100Km_). A continuación se presentamos los resultados y su interpretación:

Sesgo del Estimador de la Media

Sesgo del estimador de la media: -0.0046

Este valor cercano a 0 indica que el estimador de la media es casi insesgado, lo que significa que, en promedio, el estimador de la media no se desvía significativamente del valor verdadero de la media.

Sesgo del Estimador de la Varianza

Sesgo del estimador de la varianza: -0.0039

Este valor también es cercano a 0, lo que indica que el estimador de la varianza es casi insesgado, sugiriendo que, en promedio, el estimador de la varianza no se desvía significativamente del valor verdadero de la varianza.

Eficiencia del Estimador de la Media

Eficiencia del estimador de la media (varianza): 0.0106

La varianza del estimador de la media es bastante pequeña, lo que indica que el estimador es eficiente. Un valor menor de varianza implica una mayor eficiencia del estimador.

Eficiencia del Estimador de la Varianza

Eficiencia del estimador de la varianza (varianza): 0.0076

La varianza del estimador de la varianza también es pequeña, indicando una buena eficiencia del estimador. Un valor menor de varianza implica una mayor eficiencia del estimador.

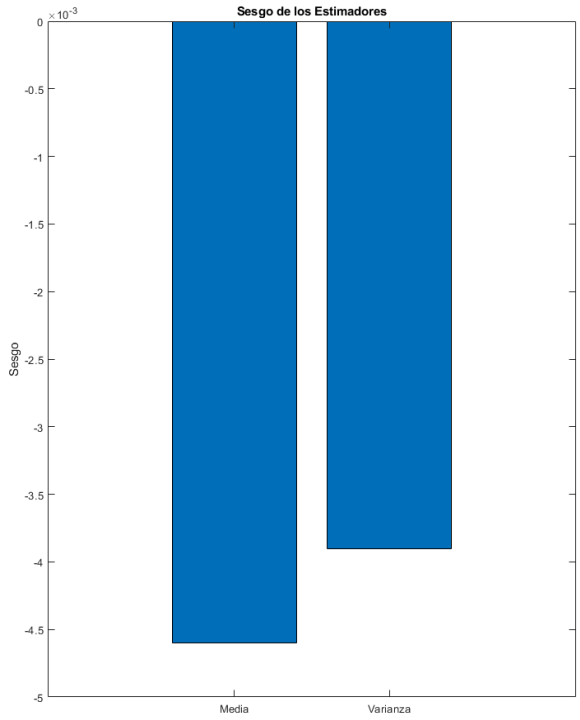
Conclusiones

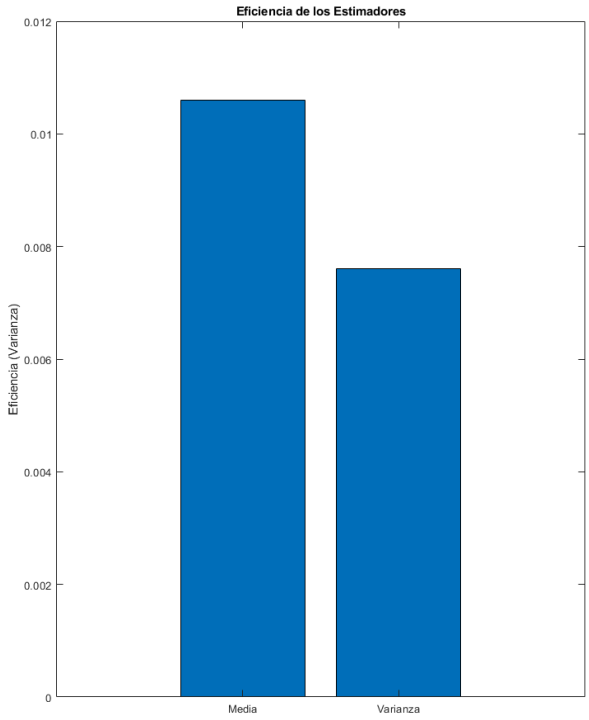
El sesgo del estimador de la media es -0.0046 y el sesgo del estimador de la varianza es -0.0039. Ambos valores son muy cercanos a 0, lo que indica que los estimadores son insesgados o casi insesgados.

Para la eficiencia vemos que la varianza del estimador de la media es 0.0106, lo que indica una alta eficiencia del estimador de la media y la varianza del estimador de la varianza es 0.0076, indicando también una alta eficiencia del estimador de la varianza.

Estos resultados sugieren que tanto el estimador de la media como el estimador de la varianza son buenos estimadores para los datos de consumo combinado de combustible. Son insesgados y tienen una alta eficiencia, lo que los hace confiables para la estimación de los parámetros verdaderos de la población.

- **Gráfica 2**





- **Proponer modelos de regresión lineal (o no lineal)**

Primero, realizamos la preparación de los datos eliminando filas con valores faltantes en la tabla F utilizando la función rmmissing. Definimos nuestra variable respuesta Y como F.Combined_L_100Km_, y seleccionamos nuestras variables predictoras: F.EngineSize_L_, F.Cylinders, F.CO2Emissions_g_km_, y las variables categóricas de F.Country utilizando la función dummyvar.

A continuación, calculamos la matriz de correlación de las variables predictoras para identificar posibles problemas de multicolinealidad. Luego, eliminamos las variables altamente correlacionadas utilizando un umbral de 0.9. Esto nos permitió reducir la multicolinealidad y mejorar la estabilidad de los coeficientes del modelo.

Con las variables reducidas, ajustamos un modelo de regresión lineal utilizando la función fitlm. Mostramos los resultados del modelo, incluyendo los coeficientes estimados, errores estándar (SE), valores t (tStat) y valores p (pValue). Estos resultados nos dieron una idea de la relación entre las variables predictoras y la variable respuesta, así como de la precisión y significancia estadística de cada coeficiente.

```
Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10

Estimated Coefficients:
                   Estimate        SE          tStat       pValue
                   ________    _________    ________    _________

    (Intercept)      0.18561      0.59015     0.31452     0.75321
    x1               0.13957     0.026225      5.3221   1.3645e-07
    x2              0.040786   0.00056438      72.267           0
    x3             -0.079061      0.55966    -0.14127      0.8877
    x4              -0.1562      0.56906    -0.27448     0.78379
    x5             -0.11971      0.55804    -0.21451     0.83021
    x6                    0            0         NaN         NaN
    x7             -0.047668      0.56226   -0.084779     0.93246
    x8              -0.1402       0.5652    -0.24806     0.80416
    x9             -0.15426      0.56144    -0.27475     0.78359
    x10            -0.051199      0.59908   -0.085463     0.93192
```

Para evaluar el desempeño del modelo de regresión lineal reducido, realizamos una validación cruzada de 10 particiones (KFold). Calculamos el error de predicción promedio (MSE) en cada partición y finalmente obtuvimos el MSE promedio para el modelo.

Resultados del Modelo de Regresión Lineal Reducido

El modelo de regresión lineal ajustado con las variables predictoras reducidas mostró coeficientes estimados que indican la relación entre cada variable predictora y la variable respuesta. Los errores estándar proporcionaron una medida de la precisión de estos coeficientes, mientras que los valores t y los valores p nos ayudaron a evaluar la significancia estadística de cada coeficiente. Observamos que algunos coeficientes tienen valores p altos, lo que sugiere que podrían no ser significativamente diferentes de cero. Esto podría indicar que ciertas variables predictoras no tienen un impacto significativo en la variable respuesta.

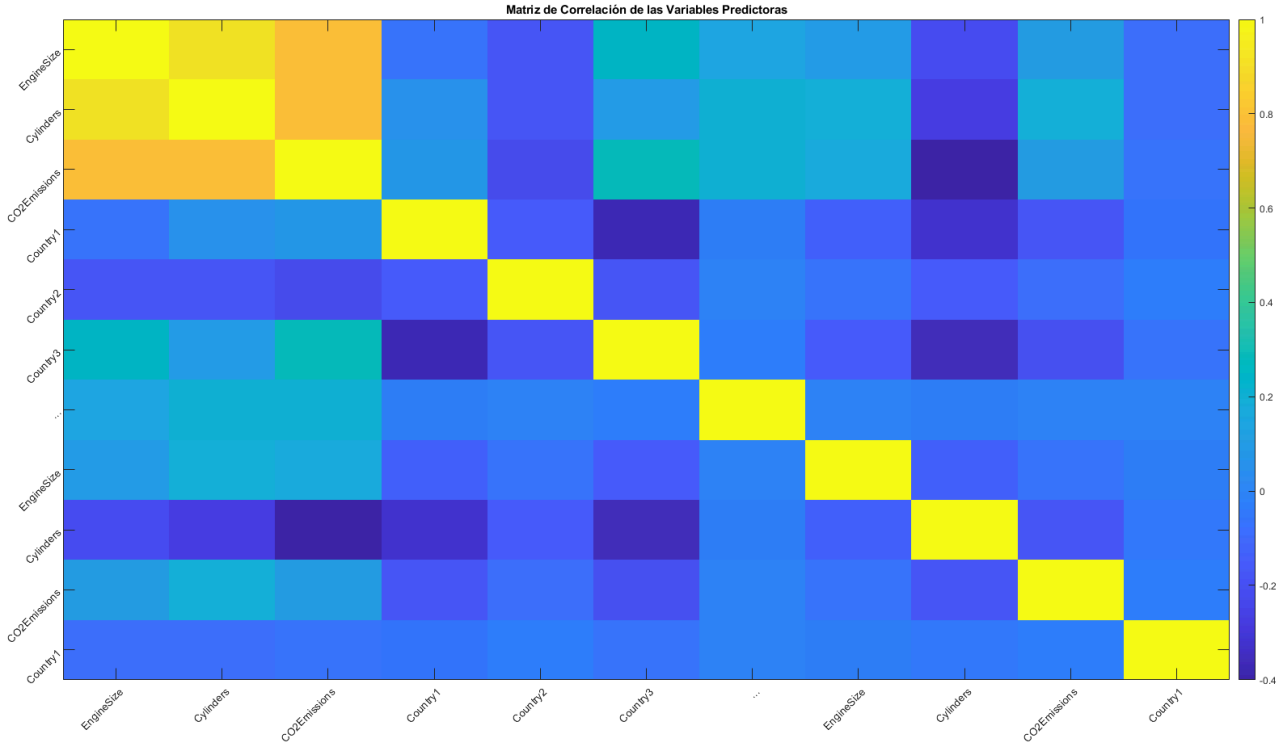
El modelo de regresión lineal con variables reducidas presentó una raíz del error cuadrático medio (RMSE) de 0.544 y un R-cuadrado ajustado de 0.964, lo que indica que el modelo explica bien la variabilidad en los datos. Además, el error de predicción promedio (MSE) obtenido mediante validación cruzada fue de 0.2978, lo que muestra que el modelo tiene un buen desempeño en la predicción de nuevos datos.

Conclusiones

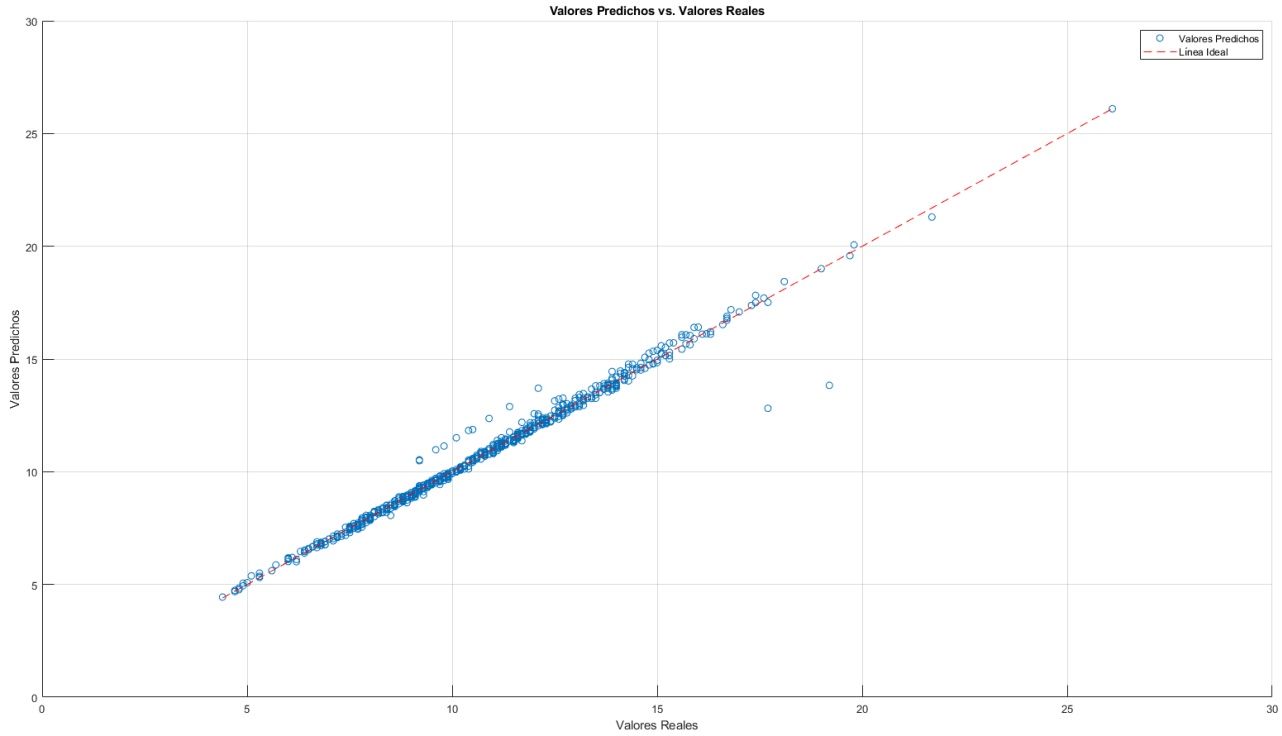
El análisis realizado nos permitió identificar y eliminar variables altamente correlacionadas, reduciendo así la multicolinealidad y mejorando la estabilidad del modelo de regresión lineal. El modelo resultante tiene un buen ajuste y un error de predicción razonable, aunque algunos coeficientes pueden no ser significativos. Estos hallazgos destacan la importancia de la selección cuidadosa de variables predictoras para construir modelos de regresión robustos y precisos.

En resumen, hemos logrado un modelo de regresión lineal que explica adecuadamente la variabilidad en el consumo combinado de combustible y que predice bien los datos nuevos. No obstante, es importante considerar la significancia estadística de cada coeficiente para asegurar que todas las variables predictoras incluidas en el modelo realmente contribuyen de manera significativa a explicar la variable respuesta.

- **Gráfica 3**



- **Gráfica 4**



- **Intervalos de confianza**

Para calcular los intervalos de confianza para la media y la varianza del consumo combinado de combustible (Combined_L_100Km_), utilizaremos métodos teóricos basados en la normalidad de los datos y los compararemos con los intervalos obtenidos mediante el método bootstrap.

Intervalo de Confianza para la Media

El intervalo de confianza para la media, asumiendo que los datos son normalmente distribuidos, se calcula como:

$$(\overline{X} - z_{a/2} \cdot \frac{\sigma}{\sqrt{n}}; \overline{X} + z_{a/2} \cdot \frac{\sigma}{\sqrt{n}})$$

Intervalo de Confianza para la Varianza

El intervalo de confianza para la varianza, asumiendo que los datos son normalmente distribuidos, se calcula utilizando la distribución chi-cuadrado:

$$(\frac{(n-1)s^2}{X^2_{a/2}}, \frac{(n-1)s^2}{X^2_{-a/2}})$$

Intervalo de Confianza para la Media

- Intervalo Teórico: El intervalo de confianza teórico para la media es [10.8449, 11.2559]. Este intervalo se basa en la suposición de que los datos siguen una distribución normal y utiliza el valor crítico de la distribución normal estándar para calcular los límites del intervalo.

- Intervalo Bootstrap: El intervalo de confianza bootstrap para la media es [10.8464, 11.2449]. Este intervalo no asume normalidad en los datos y se basa en el remuestreo repetido de los datos originales para estimar los límites del intervalo.

Comparando ambos intervalos, podemos ver que son bastante similares, lo que sugiere que la suposición de normalidad es razonable para estos datos. El intervalo bootstrap proporciona una verificación robusta del intervalo teórico, indicando que la estimación de la media es confiable.

Intervalo de Confianza para la Varianza

- Intervalo Teórico: El intervalo de confianza teórico para la varianza es [7.4052, 9.0765]. Este intervalo se calcula utilizando la distribución chi-cuadrado, asumiendo que los datos son normalmente distribuidos.

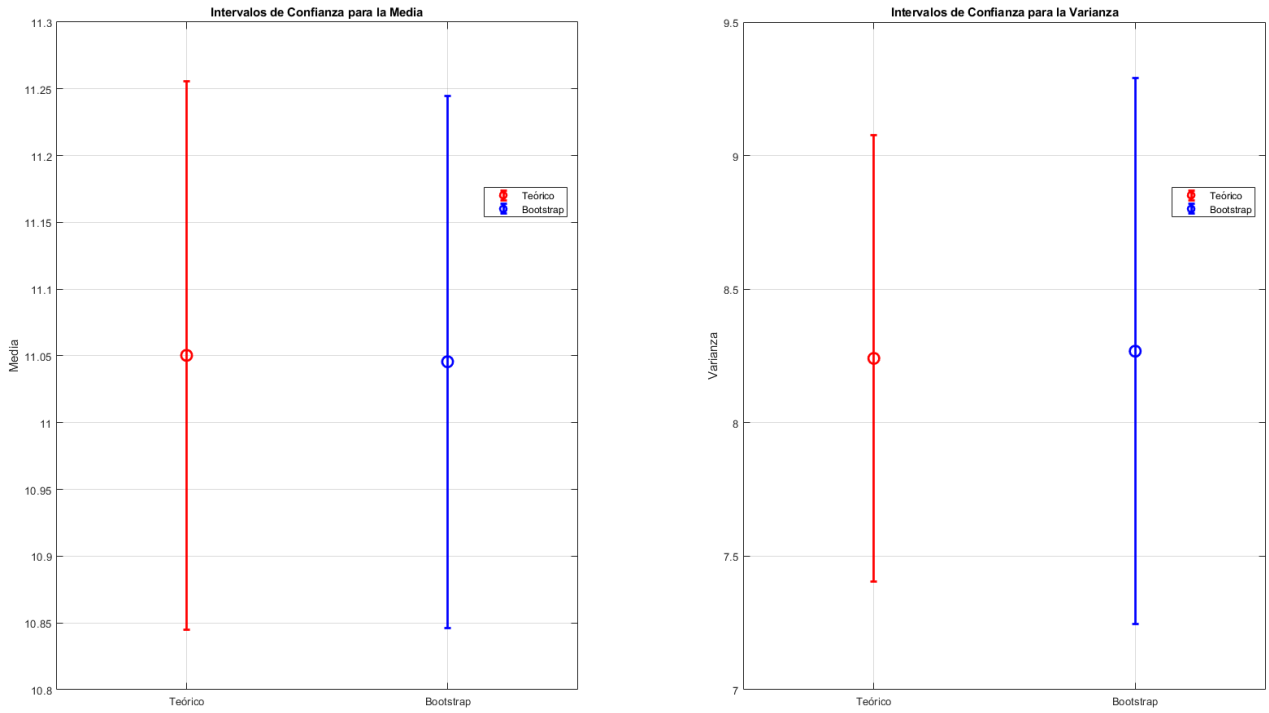
- Intervalo Bootstrap: El intervalo de confianza bootstrap para la varianza es [7.2460, 9.2894]. Similar al intervalo para la media, este método no asume normalidad y se basa en el remuestreo de los datos para estimar los límites del intervalo.

Al comparar estos intervalos, observamos que también son bastante cercanos entre sí. Esto sugiere que la suposición de normalidad para los datos de consumo combinado de combustible es adecuada y que el método bootstrap confirma la validez del intervalo teórico para la varianza.

Conclusión General

Los intervalos de confianza calculados mediante métodos teóricos y bootstrap para la media y la varianza del consumo combinado de combustible son consistentes entre sí. Esto indica que la suposición de normalidad de los datos es razonable y que nuestras estimaciones de la media y la varianza son confiables. El método bootstrap ha servido como una herramienta robusta para verificar los resultados obtenidos teóricamente, proporcionando confianza adicional en nuestras conclusiones. En resumen, hemos validado que tanto la media como la varianza del consumo combinado de combustible pueden ser estimadas de manera precisa y confiable con los métodos utilizados.

- **Gráfica 5**



- **Contraste de Hipótesis**

En este análisis, hemos llevado a cabo dos contrastes de hipótesis: uno para evaluar si la media del consumo combinado de combustible es significativamente diferente de 11 L/100 km y otro para evaluar si la varianza es significativamente diferente de 8 L/100 km. A continuación, se detallan los pasos y los resultados obtenidos.

Contraste de Hipótesis para la Media del Consumo Combinado de Combustible

- H(0): La media del consumo combinado de combustible es igual a 11 L/100 km.
- H(1): La media del consumo combinado de combustible es diferente de 11 L/100 km.

Cálculo del Estadístico de Prueba y p-valor:

Utilizamos una prueba t para una muestra para evaluar esta hipótesis.

- Estadístico t: 0.4088
- p-valor: 0.6308
- Intervalo de confianza: [10.8446, 11.2562]
- No rechaza H(0)

Interpretación:

Con un p-valor de 0.6308, que es mayor que el nivel de significancia usual de 0.05, no rechazamos la hipótesis nula. Esto sugiere que no hay evidencia suficiente para afirmar que la media del consumo combinado de combustible se desvía significativamente de 11 L/100 km.

Contraste de Hipótesis para la Varianza del Consumo Combinado de Combustible

- H(0): La varianza del consumo combinado de combustible es igual a 8 L/100 km.
- H(1):La varianza del consumo combinado de combustible es diferente de 8 L/100 km.

Cálculo del Estadístico de Prueba y p-valor:

Utilizamos una prueba chi-cuadrado para evaluar esta hipótesis.

- Estadístico chi-cuadrado: 759.4275
- p-valor: 0.6597

Interpretación:

Con un p-valor de 0.6597, que es mayor que el nivel de significancia usual de 0.05, no rechazamos la hipótesis nula. Esto indica que no hay evidencia suficiente para afirmar que la varianza del consumo combinado de combustible se desvía significativamente de 8 L/100 km.

Conclusión General

Los resultados de los contrastes de hipótesis muestran que no hay evidencia suficiente para afirmar que la media y la varianza del consumo combinado de combustible se desvían de los valores hipotéticos de 11 L/100 km y 8 L/100 km, respectivamente. Esto sugiere que los valores hipotéticos son razonables y que las estimaciones basadas en los datos muestrales no presentan desviaciones significativas. Estos hallazgos proporcionan una base sólida para entender mejor las características del consumo de combustible en los vehículos analizados.

ANEXO

```
%Distribuciones y contrastes de bondad de ajuste

combined_consumption = F.Combined_L_100Km_;

combined_consumption = combined_consumption(~isnan(combined_consumption));

mu = mean(combined_consumption);

sigma = std(combined_consumption);

fprintf('Parámetros estimados para la distribución normal:\n');

fprintf('Media (mu): %.4f\n', mu);

fprintf('Desviación estándar (sigma): %.4f\n', sigma);

paramGamma = gamfit(combined_consumption);

shape = paramGamma(1);

scale = paramGamma(2);

fprintf('Parámetros estimados para la distribución gamma:\n');

fprintf('Forma (shape): %.4f\n', shape);

fprintf('Escala (scale): %.4f\n', scale);

num_bins = 10;

[observed_freq, bin_edges] = histcounts(combined_consumption, num_bins);

expected_freq_normal = zeros(1, num_bins);

for i = 1:num_bins

   p = normcdf(bin_edges(i + 1), mu, sigma) - normcdf(bin_edges(i), mu,
sigma);

   expected_freq_normal(i) = p * length(combined_consumption);

end

expected_freq_gamma = zeros(1, num_bins);

for i = 1:num_bins

   p = gamcdf(bin_edges(i + 1), shape, scale) - gamcdf(bin_edges(i),
shape, scale);

   expected_freq_gamma(i) = p * length(combined_consumption);

end

disp('Frecuencias Observadas:');

disp(observed_freq);
```

```
disp('Frecuencias Esperadas (Normal):');

disp(expected_freq_normal);

disp('Frecuencias Esperadas (Gamma):');

disp(expected_freq_gamma);

chi2_stat_normal = sum((observed_freq - expected_freq_normal).^2 ./
expected_freq_normal);

df_normal = num_bins - 1 - 2;

p_value_normal = 1 - chi2cdf(chi2_stat_normal, df_normal);

fprintf('Resultados del contraste Chi-cuadrado para la distribución
normal:\n');

fprintf('Chi2 Estadístico: %.2f\n', chi2_stat_normal);

fprintf('Grados de Libertad: %d\n', df_normal);

fprintf('p-valor: %.4f\n', p_value_normal);

chi2_stat_gamma = sum((observed_freq - expected_freq_gamma).^2 ./
expected_freq_gamma);

df_gamma = num_bins - 1 - 2;

p_value_gamma = 1 - chi2cdf(chi2_stat_gamma, df_gamma);

fprintf('Resultados del contraste Chi-cuadrado para la distribución
gamma:\n');

fprintf('Chi2 Estadístico: %.2f\n', chi2_stat_gamma);

fprintf('Grados de Libertad: %d\n', df_gamma);

fprintf('p-valor: %.4f\n', p_value_gamma);




%Estimación paramétrica

combined_consumption = F.Combined_L_100Km_;

combined_consumption = combined_consumption(~isnan(combined_consumption));

mu_true = mean(combined_consumption);

sigma_true = std(combined_consumption);

nSim = 1000;

n = length(combined_consumption);
```

```
mu_sim = zeros(nSim, 1);

sigma_sim = zeros(nSim, 1);

rng('default');

for i = 1:nSim

   sample = datasample(combined_consumption, n);

   mu_sim(i) = mean(sample);

   sigma_sim(i) = std(sample);

end

sesgo_mu = mean(mu_sim) - mu_true;

sesgo_sigma = mean(sigma_sim) - sigma_true;

eficiencia_mu = var(mu_sim);

eficiencia_sigma = var(sigma_sim);

fprintf('Sesgo del estimador de la media: %.4f\n', sesgo_mu);

fprintf('Sesgo del estimador de la varianza: %.4f\n', sesgo_sigma);

fprintf('Eficiencia del estimador de la media (varianza): %.4f\n',
eficiencia_mu);

fprintf('Eficiencia del estimador de la varianza (varianza): %.4f\n',
eficiencia_sigma);

%Proponer modelos de regresión lineal (o no lineal)

X = [F.EngineSize_L_, F.Cylinders, F.CO2Emissions_g_km_,
dummyvar(categorical(F.Country))];

corrMatrix = corr(X);

disp(corrMatrix);

threshold = 0.9;

toRemove = [];

for i = 1:size(corrMatrix, 1)

   for j = i+1:size(corrMatrix, 2)

       if abs(corrMatrix(i, j)) > threshold

           toRemove = [toRemove, j];

       end

   end

end
```

```
X_reduced = X;

X_reduced(:, unique(toRemove)) = [];

Y = F.Combined_L_100Km_;

mdl_reduced = fitlm(X_reduced, Y);

disp(mdl_reduced);

%Intervalos de confianza

data = F.Combined_L_100Km_;

alpha = 0.05;

n = length(data);

mu = mean(data);

sigma = std(data);

z = norminv(1 - alpha/2);

CI_mu_teorico = [mu - z * sigma / sqrt(n), mu + z * sigma / sqrt(n)];

chi2_lower = chi2inv(alpha/2, n - 1);

chi2_upper = chi2inv(1 - alpha/2, n - 1);

CI_var_teorico = [(n-1) * sigma^2 / chi2_upper, (n-1) * sigma^2 /
chi2_lower];

nBoot = 1000;

bootstat_mu = bootstrp(nBoot, @mean, data);

CI_mu_bootstrap = prctile(bootstat_mu, [2.5, 97.5]);

bootstat_var = bootstrp(nBoot, @var, data);

CI_var_bootstrap = prctile(bootstat_var, [2.5, 97.5]);

fprintf('Intervalo de confianza teórico para la media: [%.4f, %.4f]\n',
CI_mu_teorico);

fprintf('Intervalo de confianza Bootstrap para la media: [%.4f, %.4f]\n',
CI_mu_bootstrap);

fprintf('Intervalo de confianza teórico para la varianza: [%.4f, %.4f]\n',
CI_var_teorico);

fprintf('Intervalo de confianza Bootstrap para la varianza: [%.4f,
%.4f]\n', CI_var_bootstrap);

%Contraste de hipotesis

combined_consumption = F.Combined_L_100Km_;
```

```
combined_consumption = combined_consumption(~isnan(combined_consumption));

mu_hipotetico = 11;

[h, p, ci, stats] = ttest(combined_consumption, mu_hipotetico);

fprintf('Resultados de la prueba t para una muestra:\n');

fprintf('Estadístico t: %.4f\n', stats.tstat);

fprintf('p-valor: %.4f\n', p);

fprintf('Intervalo de confianza: [%.4f, %.4f]\n', ci);

fprintf('Rechazar H0: %d\n', h);

var_hipotetica = 8;

n = length(combined_consumption);

var_muestral = var(combined_consumption);

chi2_stat = (n - 1) * var_muestral / var_hipotetica;

p_valor_chi2 = 2 * min(chi2cdf(chi2_stat, n - 1), 1 - chi2cdf(chi2_stat, n
- 1));

fprintf('Estadístico Chi-cuadrado: %.4f\n', chi2_stat);

fprintf('p-valor: %.4f\n', p_valor_chi2);

%Funcion de densidad no parametrica de variable numerica

combined_consumption = F.Combined_L_100Km_;

combined_consumption = combined_consumption(~isnan(combined_consumption));

[f, xi] = ksdensity(combined_consumption);

figure;

plot(xi, f, 'LineWidth', 2);

xlabel('Consumo Combinado (L/100 km)');

ylabel('Densidad de Probabilidad');

title('Estimación de la Densidad de Probabilidad mediante KDE para el
Consumo Combinado');

grid on
```