Universidad Europea de Madrid

Inferencia Estadística

2024

Francisco Guerrero Gallardo

Índice

1. Introducción

En este informe se tiene como objetivo analizar el conjunto de datos sobre las calificaciones de consumo de combustible de vehículos ligeros de 2023, proporcionado por Natural Resources Canadá. Este dataset incluye información específica del modelo sobre las calificaciones de consumo de combustible y las emisiones estimadas de dióxido de carbono para vehículos nuevos disponibles en el mercado minorista de Canadá. Los datos abarcan una amplia gama de variables como el tamaño del motor, la clase del vehículo, el tipo de transmisión, el tipo de combustible y otros parámetros relevantes. Este análisis busca proporcionar una visión detallada del rendimiento de combustible de estos vehículos, identificando patrones y tendencias significativas.

2. Estadística Descriptiva

2.1. Exploración de los datos

El primer análisis estadístico que realizaremos será la extracción de los descriptivos básicos totales de todas las variables continuas, la media, mediana, varianza, desviación típica, curtosis, coeficiente de asimetría y cuartiles primero y tercero.

Datos Numéricos

Estadísticas descriptivas

	μ	σ	Mínimo	25%	50%	75%	Máximo	Kurtosis	Asimetría
Model year	2023.00	0.00	2023.00	2023.00	2023.00	2023.00	2023.00	0.00	0.00
Engine size (L)	3.15	1.35	1.20	2.00	3.00	3.60	8.00	0.34	1.03
Cylinders	5.63	1.97	3.00	4.00	6.00	6.00	16.00	2.78	1.30
City (L/100 km)	12.43	3.46	4.40	10.10	12.10	14.60	30.30	1.67	0.69
Highway (L/100 km)	9.35	2.30	4.40	7.70	9.10	10.70	20.90	1.63	0.88
Combined (L/100 km)	11.05	2.88	4.40	9.00	10.70	12.90	26.10	1.69	0.73
Combined (mpg)	27.38	7.56	11.00	22.00	26.00	31.00	64.00	2.68	1.21
CO2 emissions (g/km)	257.47	64.26	104.00	211.00	254.00	299.00	608.00	1.96	0.64
CO2 rating	4.52	1.28	1.00	4.00	5.00	5.00	9.00	0.42	0.02
Smog rating	5.24	1.67	1.00	5.00	5.00	7.00	8.00	0.14	-0.86

2.2. Emisiones de CO2 Y Tipos de Combustible

Tipos de combustible por país

X -> Gasolina Regular
Z -> Gasolina Premium
D -> Diesel
E -> E85
B -> Electricidad
N -> Gás Natural

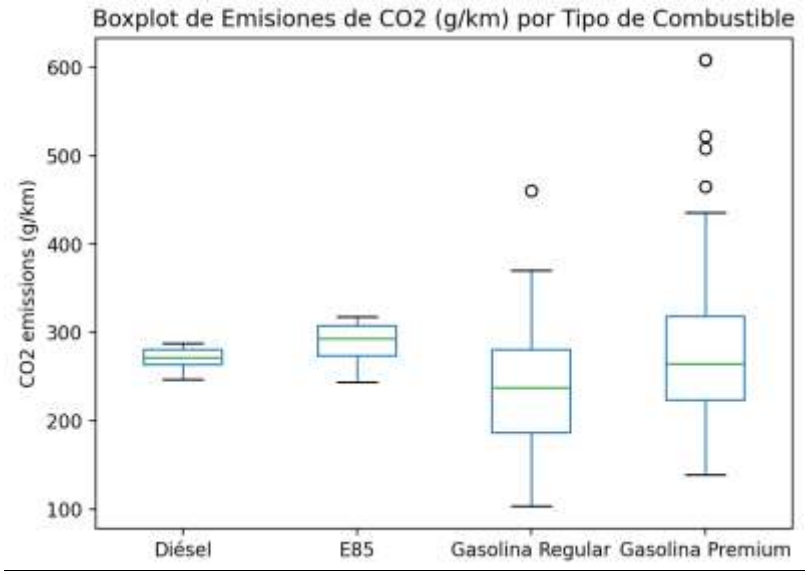
Fuente

Frecuencias de Combustible por País

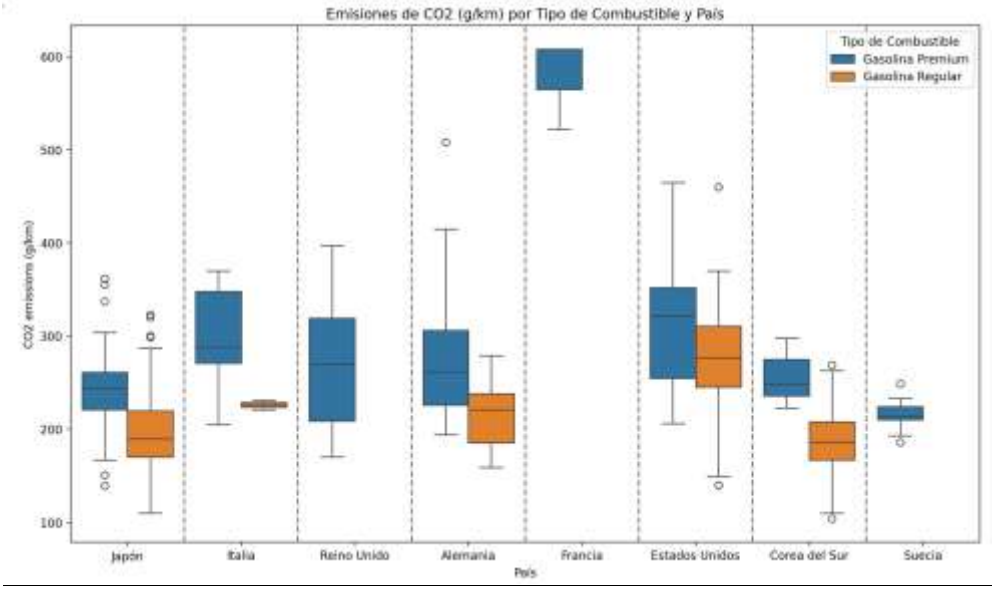
País	Gasolina Regular	Gasolina Premium	Diesel	E85	Total
Alemania	22	152	0	0	174
Corea del Sur	42	13	0	0	55
Estados Unidos	198	62	20	15	295
Francia	0	3	0	0	3
Italia	2	29	0	0	31
Japón	126	56	0	0	182
Reino Unido	0	81	0	0	81
Suecia	0	12	0	0	12
Total	**390**	**408**	**20**	**15**	**833**

En la muestra apenas aparecen vehículos que utilicen diésel o E85. La mayoría de los vehículos utilizan gasolina regular o premium y tienen una frecuencia similar (390 y 408), por lo que solo trabajaremos con estos dos tipos de combustible.

Emisiones de CO2 por cada tipo de combustible



Emisiones de CO2 por tipo de combustible y país



En este gráfico se puede apreciar como los vehículos con gasolina premium emiten ligeramente más CO2 que los vehículos con gasolina regular. Se puede apreciar un valor atípico en los vehículos con Gasolina Premium en Francia que emiten considerablemente más CO2 que el resto de los vehículos.

	Make	Model	Fuel type	City (L/100 km)	Highway (L/100 km)	Combined (L/100 km)	CO2 emissions (g/km)
128	Bugatti	Chiron	Z	26.8	16.6	22.2	522
129	Bugatti	Chiron Pur Sport	Z	30.3	20.9	26.1	608
130	Bugatti	Chiron Super Sport	Z	30.3	20.9	26.1	608

En Francia solo hay 3 vehículos que corresponden a vehículos de muy alta gama, y que presentan un consumo y unas emisiones de CO2 muy superiores al resto de vehículos.
Esto explica el valor atípico en el gráfico anterior.

3. Inferencia Estadística

Datos de la muestra

Vamos a evaluar la influencia del tipo de combustible sobre otras variables como las emisiones de CO2 o el consumo de combustible en ciudad y en autopista.

Tamaño de muestra por tipo de combustible

Gasolina Premium (Z)	408 elementos
Gasolina Regular (X)	390 elementos
Diésel (D)	20 elementos
E85 (E)	15 elementos

Como podemos observar en la imagen, apenas disponemos datos de Diesel y E85 por lo que no usaremos estos en el estudio.

Comparación de medias. Intervalo de Confianza

Se ha realizado una comparación de medias de distintos los tipos de métricas para los combustibles Gasolina Regular y Gasolina Premium mediante intervalo de confianza

¿Se puede afirmar que hay diferencia en las emisiones de CO2 entre los 2 tipos de combustible?

$\mu_x - \mu_z = \mu_d$

$H0 \equiv \mu_x = \mu_z \rightarrow \mu_d = 0$

$H1 \equiv \mu_x = \mu_z \rightarrow \mu_d \neq 0$

$\alpha = 0.05$
$n = 390$
$\bar{x} = -39.8077$
$Sd = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = 83.6157$
$t_{0.025,389} = 1.9661$
$Error = t_{\alpha/2,n-1} \cdot \frac{Sd}{\sqrt{n}} = \pm 8.3245$
$IC = \bar{x} \pm t_{\alpha/2,n-1} \cdot \frac{Sd}{\sqrt{n}} = (-48.1322, -31.4832)$
$\mu_d = 0 \notin (-48.1322, -31.4832) \rightarrow$ *Se Rechaza H0*

Se Rechaza la hipótesis nula H0 porque 0 no está incluido en el intervalo de confianza.

Por tanto, se puede afirmar que uno de los 2 combustibles contamina más.

Métrica: CO2 emissions (g/km)

Nivel de Sig... 0.050

Comparación de Medias de **Emisiones de CO2 (g/km)** entre **Gasolina Regular (X)** y **Gasolina Premium (Z)** para 390 muestras:

H0: $\mu_X = \mu_Z \Rightarrow \mu_X - \mu_Z = \mu_d = 0$

H1: $\mu_X \neq \mu_Z \Rightarrow \mu_X - \mu_Z = \mu_d \neq 0$

Media de la diferencia entre X y Z (μ_d): -39.8077

Desviación estándar de μ_d: 83.6157

Valor crítico de t para n-1 (389) grados de libertad y alpha/2 (0.025): 1.9661

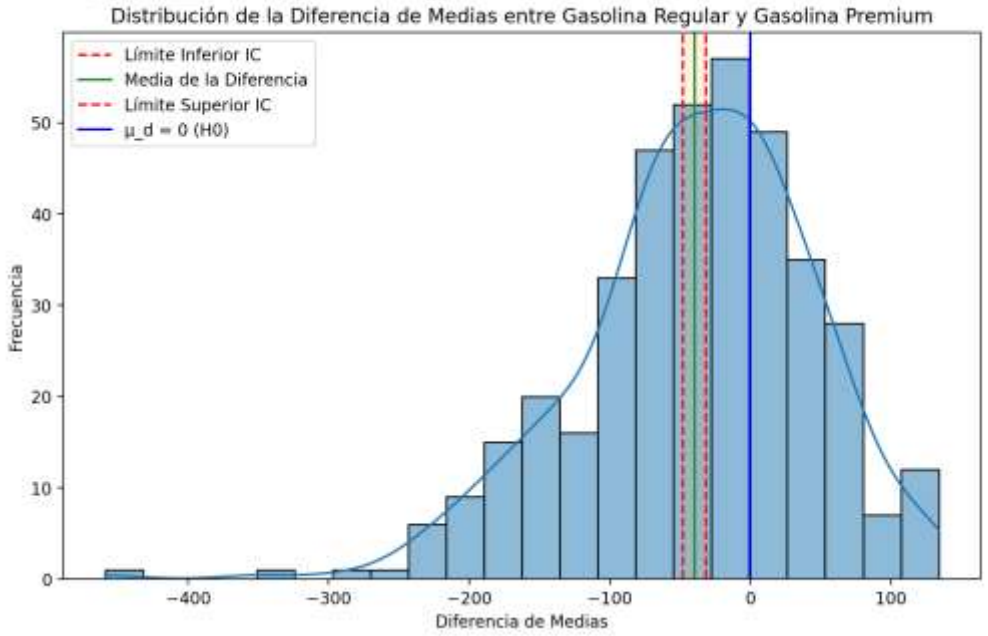
Error estándar de la diferencia entre X y Z: ±8.3245

Intervalo de Confianza (al 95%) para la diferencia de medias

IC: -48.1322 <= μ_d <= -31.4832

Resultados del Test de Hipótesis

✗ Se puede suponer que la media de Emisiones de CO2 (g/km) entre Gasolina Regular y Gasolina Premium es **diferente** ($\mu_d \neq 0$) (**se rechaza H0**)



En el código también se han hecho pruebas donde se puede comprobar que entre los 2 tipos de combustibles también hay diferencias para el consumo de combustible en ciudad y en autopista.

Contraste de Hipótesis

Vamos a realizar un nuevo contraste de hipótesis para determinar si las emisiones de CO2 son mayores en vehículos que utilizan Gasolina Premium en comparación con los que utilizan Gasolina Regular.

Planteamiento del Contraste

$H0 \equiv \mu_{Premium} \leq \mu_{Regular} \rightarrow \mu_{Premium} - \mu_{Regular} \leq 0$

$H1 \equiv \mu_{Premium} > \mu_{Regular} \rightarrow \mu_{Premium} - \mu_{Regular} > 0$

Utilizaremos un contraste de hipótesis para la diferencia de medias de dos muestras independientes con varianzas desconocidas y diferentes, utilizando la siguiente fórmula:

$$T = \frac{\bar{x}_{Premium} - \bar{x}_{Regular}}{\sqrt{\frac{Sd^2_{Premium}}{n_{Premium}} + \frac{Sd^2_{Regular}}{n_{Regular}}}}$$

$$v = \frac{\left(\frac{Sd^2_{Regular}}{n_{Regular}} + \frac{Sd^2_{Premium}}{n_{Premium}}\right)^2}{\frac{\left(\frac{Sd^2_{Regular}}{n_{Regular}}\right)^2}{n_{Regular} - 1} + \frac{\left(\frac{Sd^2_{Premium}}{n_{Premium}}\right)^2}{n_{Premium} - 1}}$$

$$IC = \left(\bar{x}_{Premium} - \bar{x}_{Regular}\right) \pm t_{\alpha,v} \cdot \sqrt{\frac{Sd^2_{Premium}}{n_{Premium}} + \frac{Sd^2_{Regular}}{n_{Regular}}}$$

$\alpha = 0.05$
$\bar{x}_{Premium} = 274.826$
$\bar{x}_{Regular} = 237.5103$
$n_{Premium} = 408$
$n_{Regular} = 390$
$Sd^2_{Premium} = 65.3299$
$Sd^2_{Regular} = 59.8201$

$T = \frac{274.83 - 237.51}{\sqrt{\frac{59.82^2}{390} + \frac{65.33^2}{408}}} = \mathbf{8.421}$

$v = \frac{\left(\frac{59.82^2}{390} + \frac{65.33^2}{408}\right)^2}{\frac{\left(\frac{59.82^2}{390}\right)^2}{390-1} + \frac{\left(\frac{65.33^2}{408}\right)^2}{408-1}} = \mathbf{794.541}$

$t_{Crítico} = t_{794.5,0.05} = \mathbf{1.6468}$

8.421 > **1.6468** → Se rechaza H0 → Se puede suponer que las emisiones de CO2 son mayores en vehículos que utilizan Gasolina Premium en comparación con los que utilizan Gasolina Regular

Conclusión

Los resultados indican que las emisiones de CO2 de los vehículos que utilizan Gasolina Premium son significativamente mayores que las de los vehículos que utilizan Gasolina Regular. Este análisis respalda el estudio anterior donde mediante intervalos de confianza, se pudo comprobar que existía una diferencia apreciable en cuanto a las emisiones de CO2 entre los dos tipos de combustibles.

Nivel de Sig... 0.050

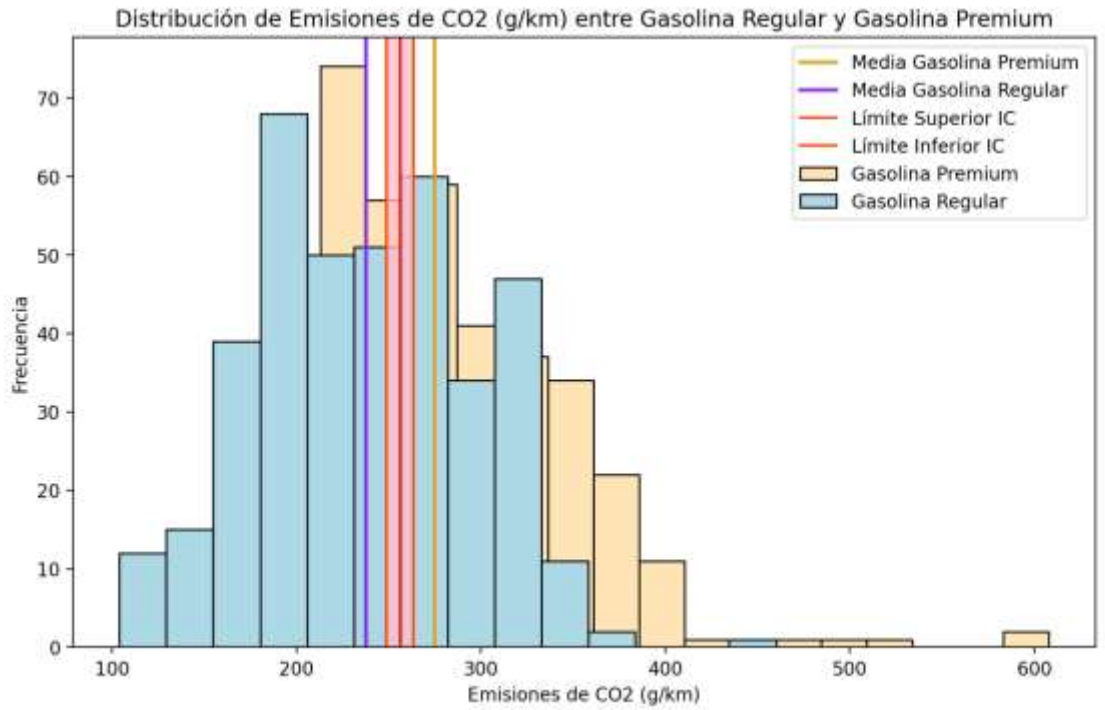
Comparación de Medias de **Emisiones de CO2 (g/km)** entre **Gasolina Premium (Z)** y **Gasolina Regular (X)**:
H0: μ_Z <= μ_X => μ_Z - μ_X <= 0
H1: μ_Z > μ_X => μ_Z - μ_X > 0

Media de Emisiones de CO2 (g/km) para Gasolina Regular (X): 237.5103
Media de Emisiones de CO2 (g/km) para Gasolina Premium (Z): 274.826
Número de muestras para Gasolina Regular (X): 390
Número de muestras para Gasolina Premium (Z): 408
Desviación estándar de Emisiones de CO2 (g/km) para Gasolina Regular (X): 59.8201
Desviación estándar de Emisiones de CO2 (g/km) para Gasolina Premium (Z): 65.3299
T: 8.421
Grados de libertad: 794.541
Valor crítico de t para un nivel de significancia de 0.05 y 794.541 grados de libertad: 1.6468

Resultados del Test de Hipótesis

✗ 8.421 > 1.6468: Se puede suponer que las emisiones de CO2 son **mayores** en vehículos que utilizan Gasolina Premium en comparación con los que utilizan Gasolina Regular (**se rechaza H0**)

Intervalo de Confianza al 95.0%: 30.0184 <= μ_Z - μ_X <= 44.613
Esto indica que, con un 95.0% de confianza, los vehículos que utilizan Gasolina Premium emiten entre 30.0184 g/km y 44.613 g/km más de CO2 en comparación con los que utilizan Gasolina Regular.



Prueba de Bondad de Ajuste

En este apartado se quiere comprobar si las emisiones de CO2 de la muestra sigue algún tipo de distribución (Normal, Gamma, Exponencial…).

En este caso, visualmente la muestra se asimila mucho a una distribución normal por lo que se procede a hacer un contraste de bondad de ajuste por el método de prueba χ^2 de Pearson.

$H0 \equiv Y \sim \mathcal{N}(257.47, 64.2243)$

$H1 \equiv \mathrm{Y} \nsim \mathcal{N}(257.47, 64.2243)$

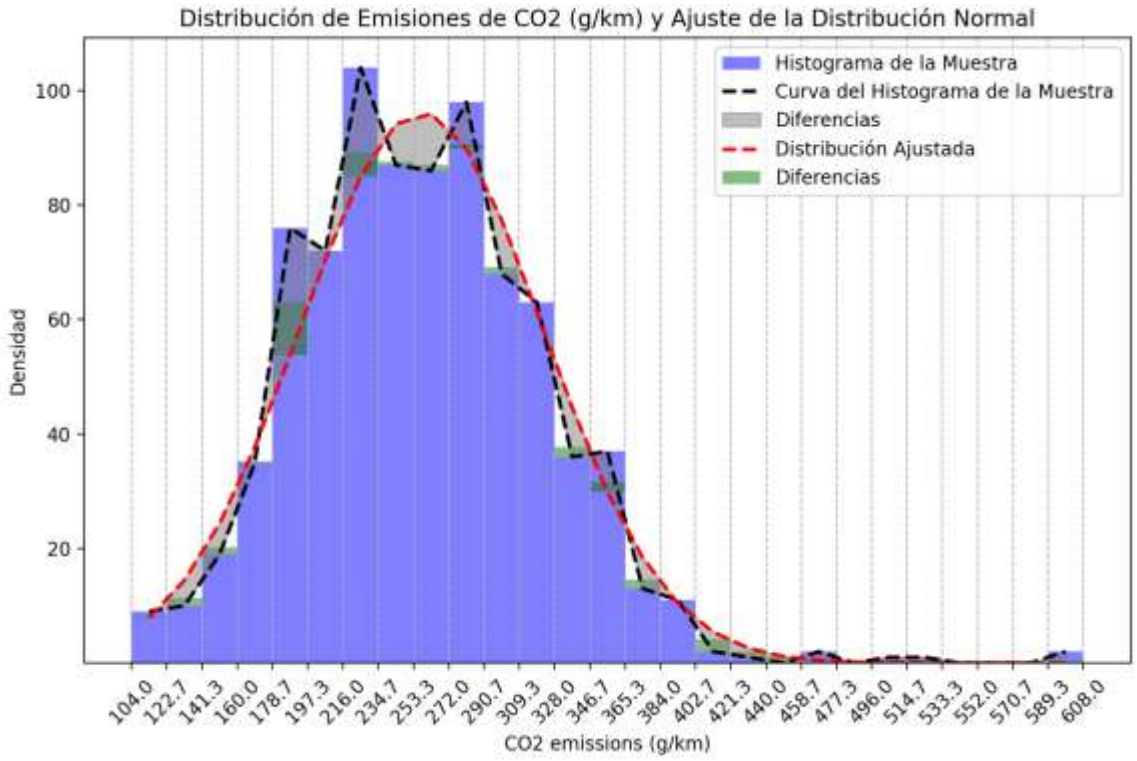
$\alpha = 0.05$
$n = 390$

s = 2 parámetros estimados
$\hat{\sigma} = 64.2243$
$\hat{\mu} = 257.47$
$O_i \sim \mathcal{N}(\mu, \sigma)$

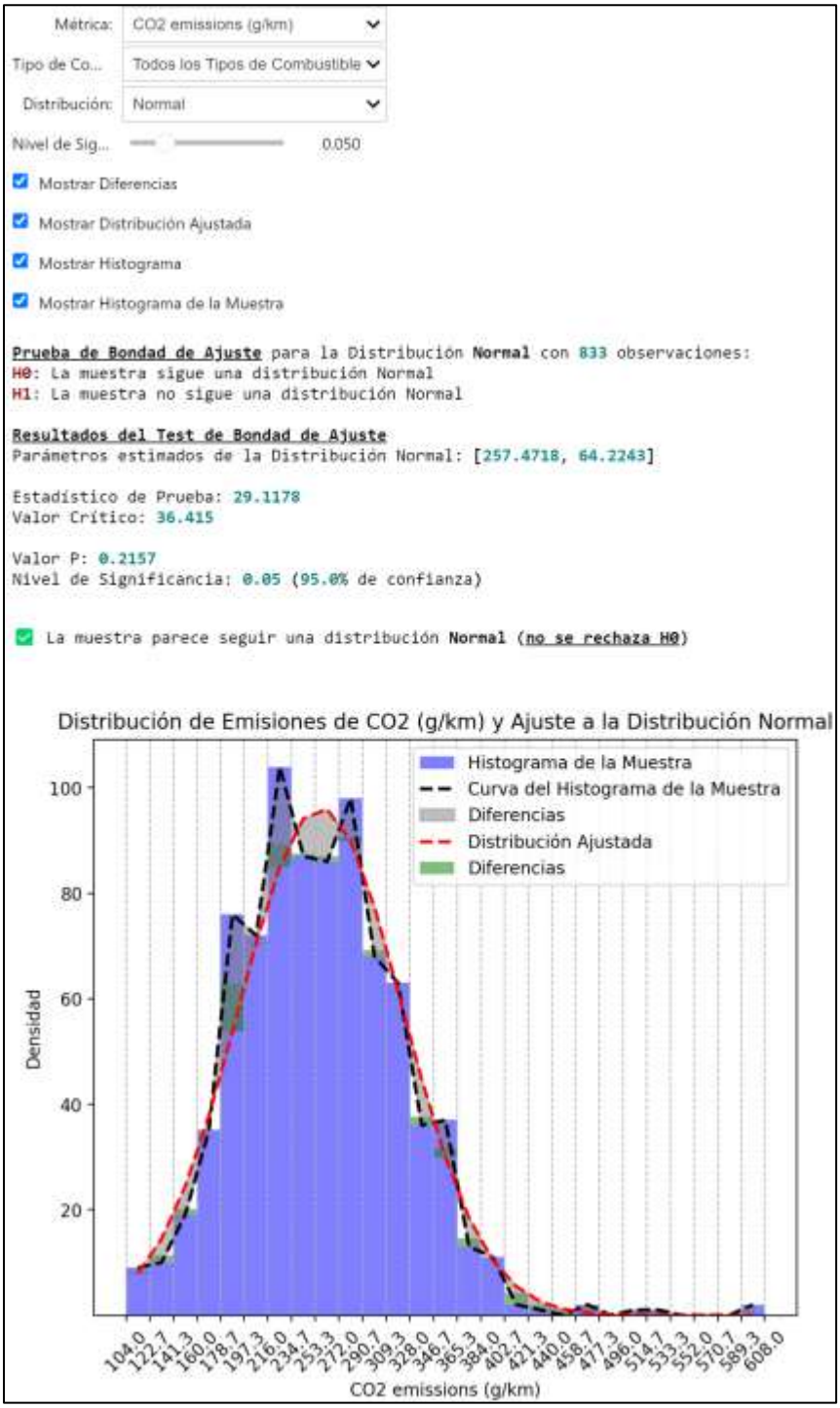
$\mathrm{T} = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{\alpha, n-1-\mathrm{s}} \equiv \mathrm{T} = 29.1178$
$\chi^2_{0.05,833-1-2} = 36.415$
$\mathrm{p-valor} = 0.2157$

$$T < \chi^2_{\alpha, n-1-\mathrm{s}} \rightarrow No\ se\ Rechaza\ H0.$$

La muestra si parece seguir una Distribución normal

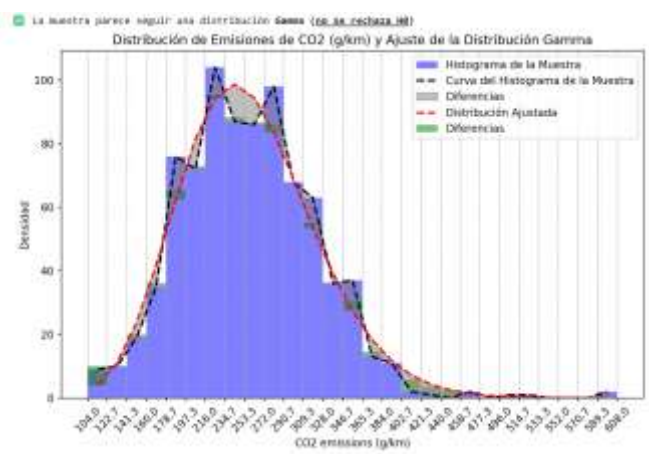


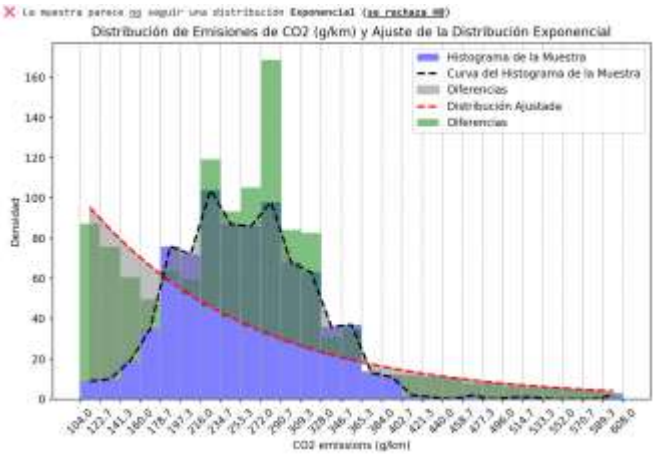
En el código se presenta una herramienta para el cálculo de la prueba de Bondad de Ajuste donde, de forma interactiva, se pueden modificar las variables y evaluar la prueba para distintas distribuciones y métricas.

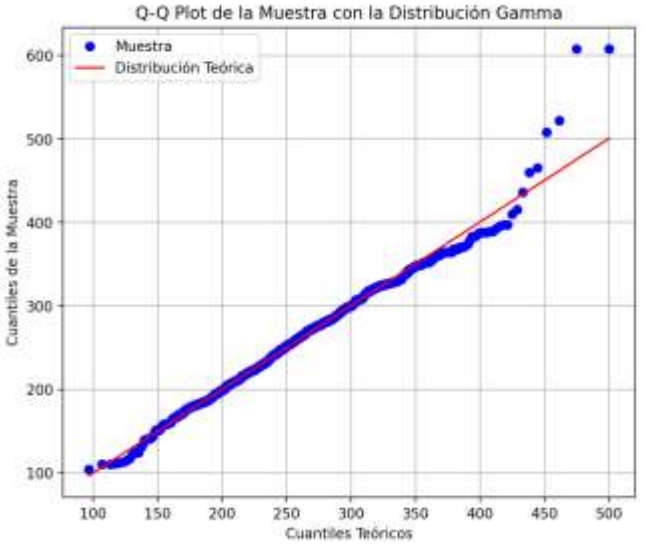


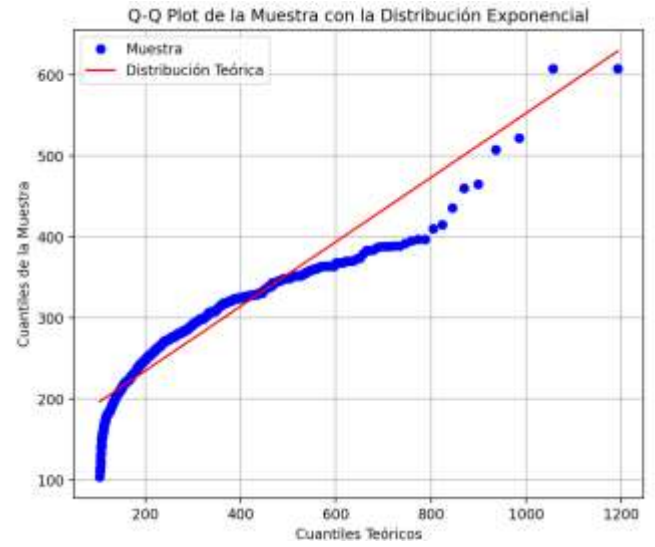
Las diferencias que aparecen en verde representan $\frac{(O_i - E_i)^2}{E_i}$ por cada bin y se puede activar su visualización activando la casilla "*Mostrar Diferencias*".

La muestra de Emisiones de CO2 en concreto también se ajusta a otras distribuciones como la Gamma y no se ajusta a la exponencial.









4. Conclusión

El tipo de combustible afecta a la cantidad de co2 emitida. En concreto la gasolina premium contamina más que la Gasolina regular.

También se ha podido observar como las emisiones de CO2 siguen una distribución normal (y también se ajusta a la distribución gamma).

La muestra es demasiado pequeña, no permitiendo sacar conclusiones sobre, por ejemplo, otros tipos de combustible.

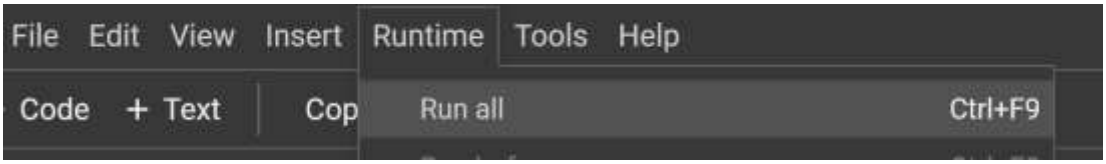
El código se encuentra en el archivo trabajo.ipynb adjuntado así como en el repositorio de GitHub:

https://github.com/xliee/inferencia-trabajo-2024.git

Este código se puede probar para poder ejecutar las herramientas interactivas desarrolladas desde Google Colab subiendo trabajo.ipynb o a través del siguiente enlace:

https://colab.research.google.com/github/xliee/inferencia-trabajo-2024/blob/main/trabajo.ipynb

Una vez dentro, debería ejecutar todas las celdas.



Una vez ejecutadas aparecerán todos los gráficos y gráficos interactivos.