

# CONTAMINACIÓN POR AUTOMÓVILES



Equipo 1: Cerecero Peña A.S., Rangel Vallejo I.A., Aguilar Sánchez J.A., García Gutiérrez A.J. y Romero Banda J.M.



Base



Código



## INTRODUCCIÓN

Los automóviles particulares generan el 18% de las emisiones totales de CO<sub>2</sub>. Una herramienta capaz de cuantificar los datos relevantes a las emisiones de CO<sub>2</sub> producidas por automóviles ayudaría a la humanidad a combatir este fenómeno que está afectando a la Tierra y su población.

## OBJETIVOS

- PRINCIPAL: Predecir las emisiones de CO<sub>2</sub> (g/km) de los automóviles.
- SECUNDARIO: Establecer si el tipo de gasolina influye en las emisiones producidas.

## PREGUNTAS DE INTERÉS

- ¿Qué clases de vehículo representan un mayor riesgo al ambiente?
- ¿La transmisión del automóvil hace una diferencia al momento de emitir CO<sub>2</sub>?

Recursos



## METODOLOGÍA

Antes de trabajar de lleno con los objetivos, se empieza con el análisis de la base de datos. En la base se tuvieron que hacer modificaciones; se cambió el nombre de las columnas con el fin de hacer mención a ellas fácilmente, por lo que sólo se agregó un guion bajo ( \_ ); también se modificó la categorización de la columna "Transmission" mediante un ciclo, para tener únicamente transmisiones automáticas (A) y manuales (M). Posterior a esto, se observó que la columna "Fuel Consumption Comb (mpg)" mantiene una relación inversa con las demás (figura 4), por lo que se decidió eliminarla. Esto no supone un problema puesto que esta columna es explicada por "Fuel Consumption Comb (L/100 km)". Nótese que sólo tienen distintas unidades de medición. Después de estos pasos, se trabajó con la base resultante ya que no se tenían datos nulos.

## GRÁFICAS EMPLEADAS

Las figuras 1, 2 y 3 revelan información importante que servirá para concluir el objetivo secundario y las preguntas de interés. En cambio, la figura 4 demuestra que las columnas de la base (variables independientes) se relacionan fuertemente con las emisiones de CO<sub>2</sub> (variable dependiente), por lo que se decidió utilizar la técnica de regresión lineal para contestar el objetivo principal.

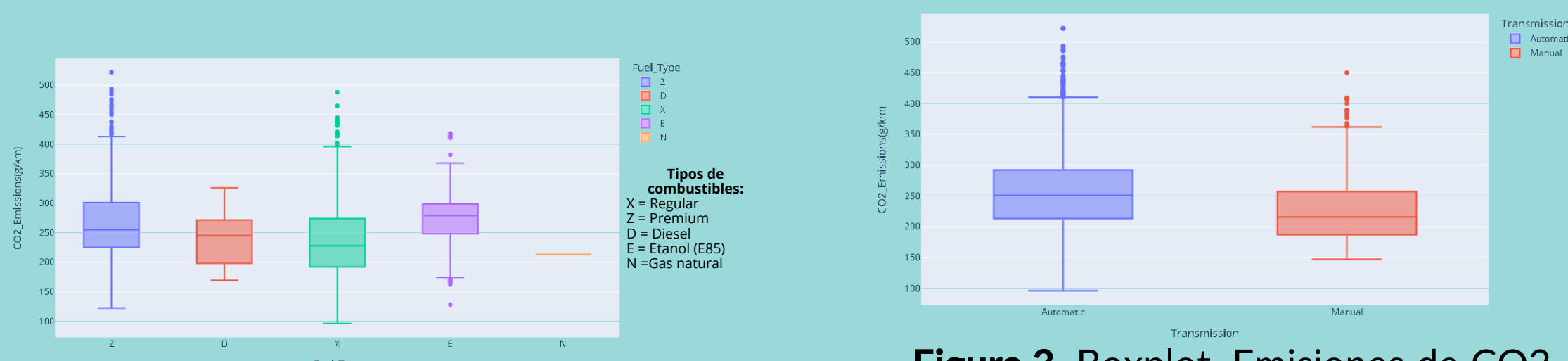


Figura 1. Boxplot. Emisiones de CO<sub>2</sub> por tipo de combustible.

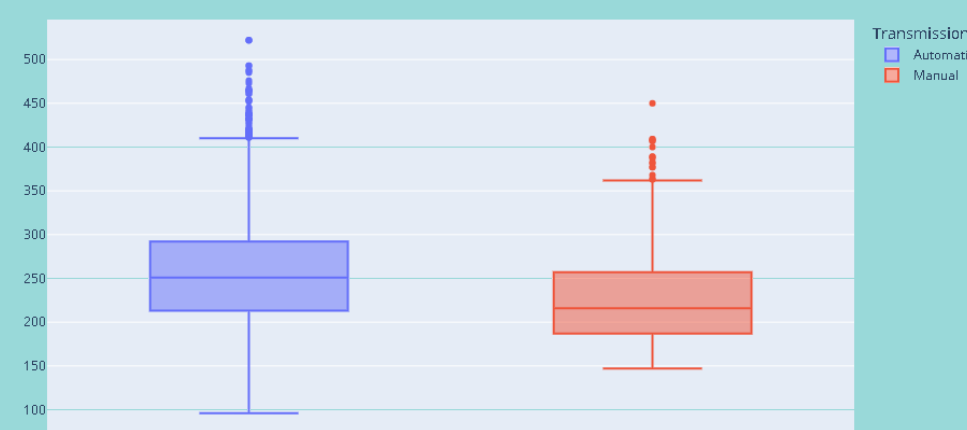


Figura 3. Boxplot. Emisiones de CO<sub>2</sub> por tipo de transmisión.

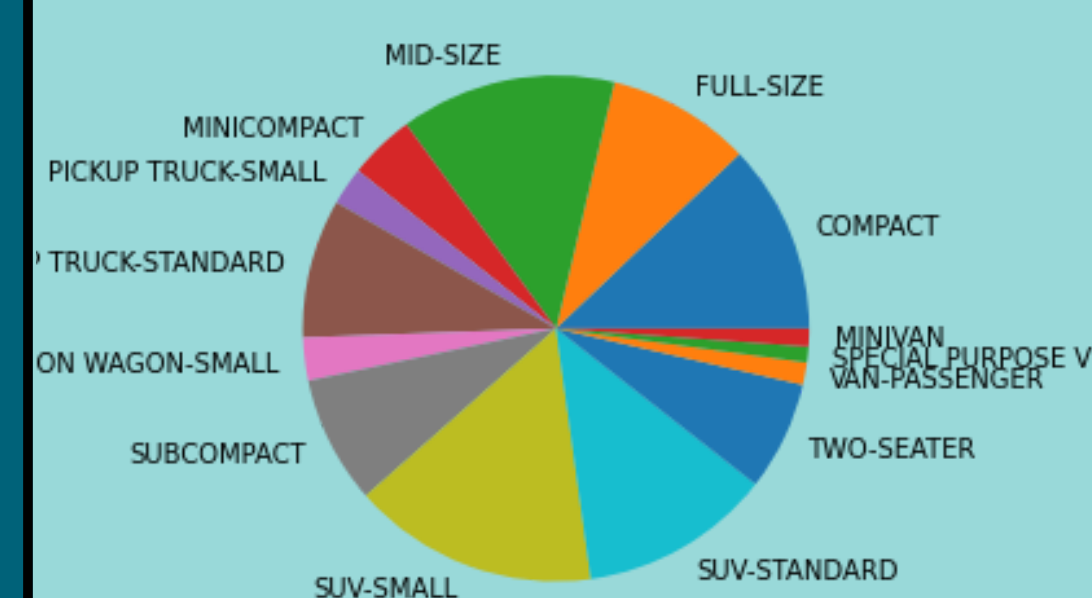


Figura 2. Gráfica de pastel. Emisiones de CO<sub>2</sub> por clase de vehículo.

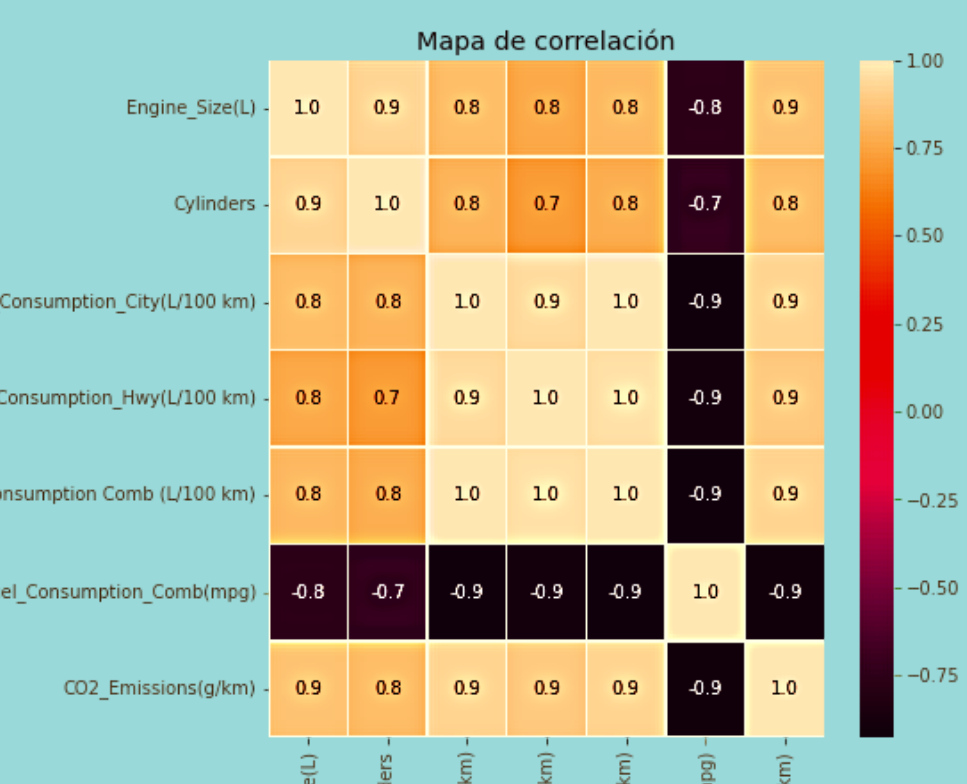


Figura 4. Matriz de correlación entre columnas de la base de datos.

## RESULTADOS

La técnica utilizada para cumplir el objetivo principal es la regresión lineal, ya que ayuda a predecir las emisiones de CO<sub>2</sub> mediante un modelo. En el mejor modelo parcial, las variables independientes son: "Engine\_Size(L)", "Cylinders" y "Fuel Consumption Comb (L/100km)"; mientras que la variable dependiente es "CO<sub>2</sub> Emissions(g/km)". Cabe aclarar que se eliminaron datos atípicos (outliers); además, del 100% de los datos conservados, se utilizó el 80% para obtener la regresión lineal múltiple y los restantes fueron aprovechados para medir la exactitud de las predicciones calculadas a partir del modelo que se muestra a continuación:

$$\text{Emisiones de CO}_2(\text{g/km}) = 55.0108 + 12.8562 \text{FuelConsumptionComb}(\text{L}/100\text{km}) + 5.8884 \text{EngineSize}(\text{L}) + 6.2895 \text{Cylinder}$$

El **ajuste** obtenido en el modelo es de **86.9%**, un muy buen ajuste dado que explica cerca del 90% de la variabilidad de los datos. Por otra parte, se calculó el valor del **error cuadrático medio** que estima, en promedio, cuánto se desvían los valores estimados de los reales, dicha desviación tiene un valor de **19.17(g/km)**.

### Cumplimiento de los supuestos de los residuales

- Media cero
- Normalidad
- Varianza const.
- Incorrelación

### Supuestos de los residuales

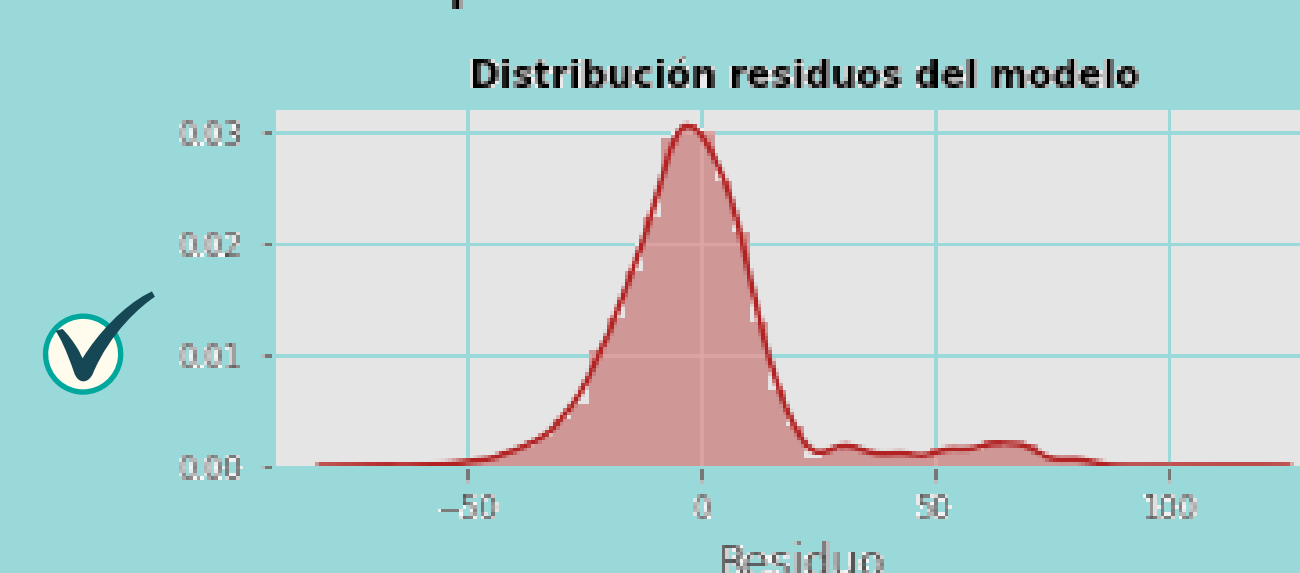


Figura 5. Distribución de los residuos del modelo. Observe que estos tienen media cero.

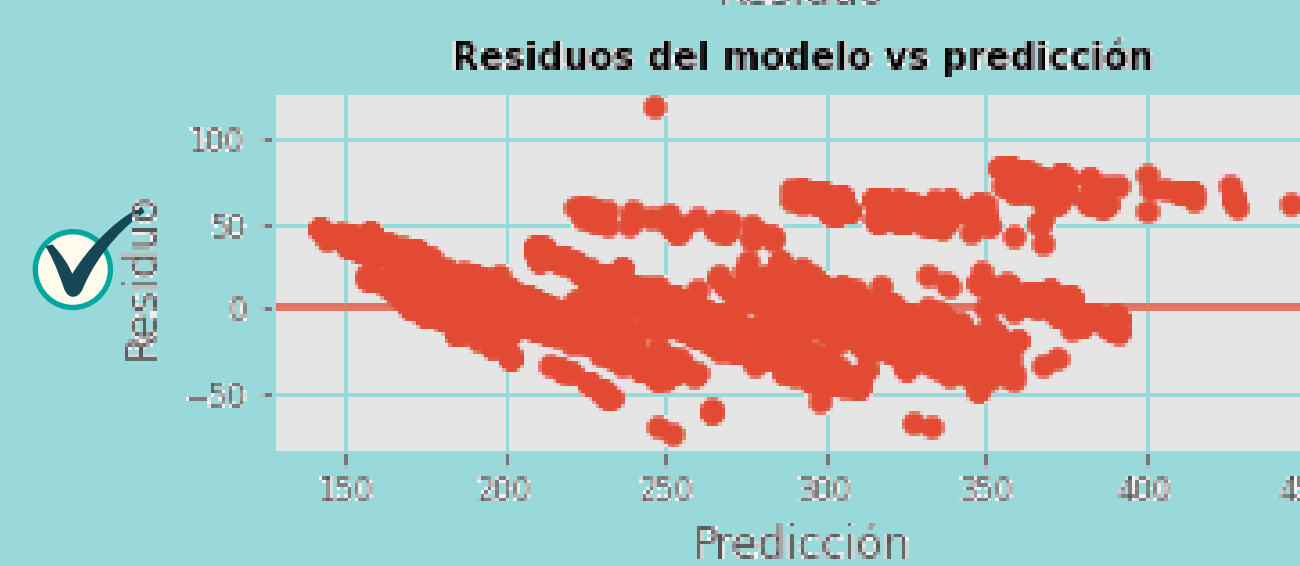


Figura 6. Dispersión de los residuos del modelo vs predicciones. Note la presencia de una varianza constante

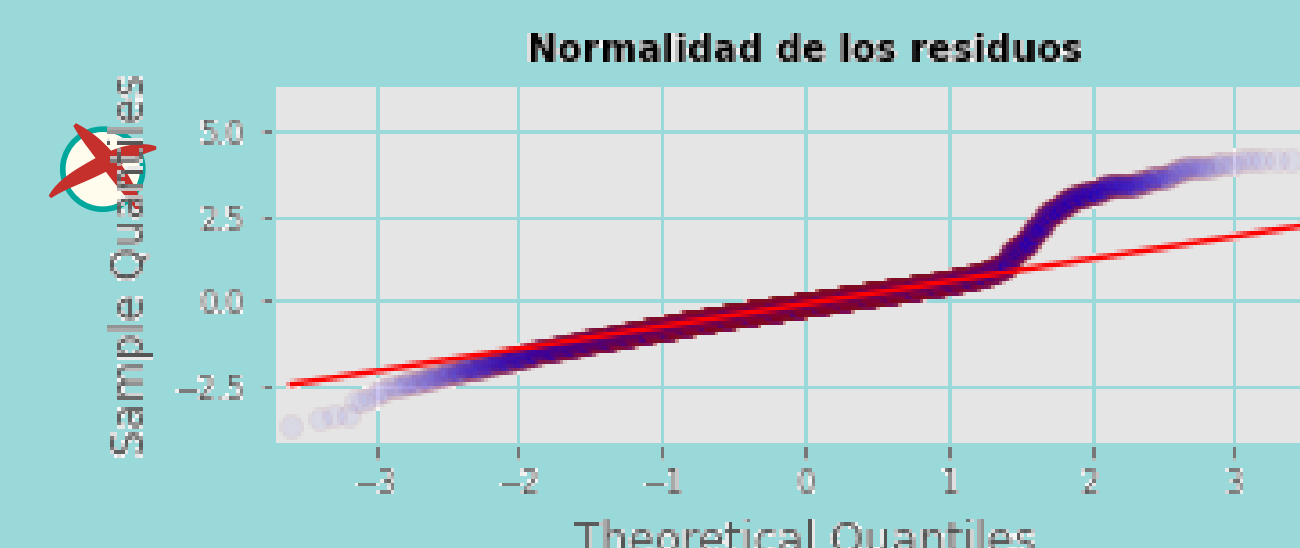


Figura 7. QQPLOT de los residuos. No cumple los valores teóricos.

## CONCLUSIONES

- A partir de la **figura 1** se concluye que existe diferencia entre usar la gasolina regular (X) y premium (Z), así como también existe entre usar etanol (E) y gas natural (N). Por tanto, el combustible influye en las emisiones que un automóvil produce.
- Basado en la **figura 2**, se llega a la conclusión de que las clases de vehículos: 'SUV-SMALL', 'SUV-STANDARD', 'MID-SIZE' y 'COMPACT' son las principales emisoras de CO<sub>2</sub>. Por ende, representan mayor riesgo al ambiente.
- De la **figura 3** se tiene que el tipo de transmisión utilizada en el vehículo influye en la emisión de CO<sub>2</sub>, ya que los automáticos afectan más al ambiente que los manuales.
- Los residuales del modelo no cumplen con todos los supuestos, esto indica que las variables utilizadas se relacionan fuertemente entre ellas; se observa en la **figura 4**. A pesar de esto, el modelo se puede utilizar para contestar el objetivo principal ya que esto afecta únicamente a los coeficientes de la ecuación (con variaciones).
- Del modelo lineal múltiple, se pueden destacar dos valores: **ajuste** (86.9%) y **error cuadrático medio** (19.17(g/km)); estas métricas permiten hacer el trabajo comparable con otro, ya que a mayor ajuste y menor error, un modelo se considera mejor. Gracias a estos valores se puede establecer qué tan bueno es el modelo obtenido. Se apreció que se pueden predecir las emisiones de CO<sub>2</sub> de un automóvil en función del tamaño del motor, el número de cilindros y la cantidad de combustible utilizado por cada 100 km; la ecuación del modelo permite predecir las emisiones de CO<sub>2</sub> siempre y cuando se tengan registros de los valores para las variables independientes.

## IDEAS PARA MEJORAR EL RESULTADO

- Agregar variables al modelo que expliquen las emisiones de CO<sub>2</sub> de los automóviles pero que no se relacionen entre ellas, puesto que esta relación causaría problemas en la ecuación del modelo de regresión lineal múltiple afectando los supuestos de los residuales.
- Probar diferentes modelos y compararlos entre ellos. Generando un modelo para cierto número de variables y determinar cuál es el mejor con el fin de no recabar más datos pero sí conocer el mejor modelo de las opciones que se tienen.

## ¿QUÉ SE PUEDE HACER CON LOS DATOS OBTENIDOS?

- Con los datos generados, es posible resolver o tomar medidas en contra de la contaminación del medio ambiente, ya que en los últimos años ha tomado más peso el tema de la contaminación producida por el uso del automóvil, abrir el tema de discusión y tratar de plantear soluciones que ayuden a remediar o disminuir el impacto generado por la utilización de estos.