Introducción a la Minería de Datos: Trabajo Datos atípicos y faltantes

Juan David Garcia Zapata[a,1], **Juan Camilo Sandoval Macías**[b,2], **Katerin Gomez Castrillon**[b,c,3]

[a]Departmento de Estadistica; [b]Universidad Nacional de Colombia; [c]Sede Medellin

Julieth Veronica Guarin Escudero

Abstract*—The preprocessing stage aims to address techniques that allow for obtaining organized information, avoiding redundant data, identifying potential issues in databases, and applying the respective treatments. Specifically, this work will develop methodologies for handling outliers and missing data. These techniques aim to optimize the quality and integrity of the data, ensuring a solid foundation for subsequent analysis.

1. Introduction

En la etapa de preprocesamiento se tiene como objetivo abordar técnicas que permitan obtener información organizada, evitar información que pueda ser redundante, identificar posibles problemas presentes en las bases de datos y hacer el respectivo tratamiento. Específicamente, en este trabajo se desarrollarán metodologías para el tratamiento de datos atípicos y datos faltantes.

Parte A

1.1. Sobre el dataset ruidoso

El ruido se mide comúnmente en decibeles (dB), una unidad logarítmica que se utiliza para expresar la intensidad relativa del sonido. La escala de decibeles se basa en la relación entre la presión sonora medida y una referencia estándar, que es el umbral de audición humana típica en condiciones específicas.

El sonido se percibe de manera no lineal por el oído humano, por lo que la escala de decibeles se diseñó para reflejar esta percepción. Algunos puntos importantes sobre la escala de decibeles son:

- **Umbral de audición**: Se define como 0 dB y representa el nivel de sonido más débil que un oído humano promedio puede percibir. Este valor se toma como referencia.
- **Nivel de sonido aumentado**: Un incremento de 10 dB representa aproximadamente una duplicación en la percepción del sonido.
- **Nivel de sonido disminuido**: Un decrecimiento de 10 dB representa aproximadamente una disminución a la mitad en la percepción del sonido.

La fórmula general para calcular el nivel de presión sonora en decibeles (dB) a partir de la presión sonora (P) en pascales, en comparación con la presión de referencia (Pref), se expresa como:

$$L_{dB} = 20log_{10}(\frac{P}{P_{ref}})$$

- L_{lB}: Es el nivel de presión sonora en decibeles (dB).
- **P**: Es la presión sonora medida.
- P_{ref}: Es la presión de referencia.

2. A partir del dataset ruidoso.txt realice los siguientes análisis:

2.1. Cargue y explore el dataset explicando en qué consiste y las características que posee el mismo.

Dataset:

Table 1. *Ruidoso*

Road_55dB	Road_60dB	Railways_65dB	Industry_65dB
0	166400	79200	1100
1	20000	11000	0
2	37800	18000	0
3	24500	16700	0
4	103100	33500	1900
...	...	...	...
61	20700	16200	0
62	15100	9900	0
63	24500	12400	300
64	12800	8500	0
65	3108200	1469100	25800

El conjunto de datos proporciona información sobre la exposición al ruido en varios entornos, como carreteras, ferrocarriles e industrias. Sin embargo, surgen dudas sobre la escala de medición de las variables. Aunque en (1.1) se menciona que la escala de medida es en decibeles (dB), al examinar los valores en la base de datos, se observa que estos superan el rango típico de medición en dB. Esto dificulta la comprensión de cómo se obtuvieron los datos. Se considera la posibilidad de que aún no se haya aplicado el logaritmo en base 10 a los valores. Sin embargo, al realizar esta transformación, surge otro problema: la gran mayoría de los valores en la base son 0, lo que resulta en valores NaN al aplicar el logaritmo. Por lo tanto, se opta por trabajar con los datos originales y considerar el 0 como un "cero relativo", teniendo en cuenta su significado relativo en el contexto de la medición de ruido en lugar de como una ausencia absoluta de ruido.

Las variables que representa el dataset son:

- **"Unnamed:0"**: Identificador de la observación.
- **"Road_55dB"**: La medida de exposición al ruido debido al tráfico de carreteras a 55 de decibelios.
- **"Road_60dB"**: La medida de exposición al ruido debido al tráfico de carreteras a 60 de decibelios.
- **"Railways_65dB"**: La medida de la exposición al ruido de ferrocarriles a 65 de decibelios a la que una persona está expuesta.
- **"Industry_65dB"**: la medida de la exposición al ruido industrial a 65 de decibelios.

2.2. ¿Cuantos resgistros hay?

Table 2. *Registros*

Columna	Cantidad no nula
Unnamed: 0	66
Road_55dB	66
Road_60dB	66
Railways_65dB	66
Industry_65dB	65
Total	330

El dataset contiene 330 registros, pero no todas las columnas tienen 66 registros, ya que la variable "Industry_65dB" solo cuenta con 65 registros. Esto evidencia la presencia de datos faltantes.

Datos NaN's:

Table 3. *Datos faltantes*

Columna	NaN
Unnamed: 0	0
Road_55dB	0
Road_60dB	0
Railways_65dB	0
Industry_65dB	1
Total	(330)

Se verifica la existencia de un dato faltante en la variable "Industry_65dB". Este hallazgo será abordado posteriormente dentro del marco de las prácticas estadísticas para su eventual corrección.

2.3. Tipo de datos.

Table 4. *Tipo*

Columna	Type
Unnamed: 0	Entero
Road_55dB	Entero
Road_60dB	Entero
Railways_65dB	Entero
Industry_65dB	Flotante

En resumen, el dataset está compuesto por cuatro variables enteras y una variable de punto flotante.

Eliminacion de variable

La variable "Unnamed:0" será eliminada del dataset, dado que no aporta información relevante para el análisis y solo actúa como un identificador que no tiene ningún interés en el contexto del trabajo.

2.4. Analisis inicial.

Resumen estadistico.

Table 5. *Estadísticas de los datos*

	Road_55dB	Road_60dB	Railways_65dB	Industry_65dB
Count	66.000	66.000	66.000	65.000
Mean	159228.800	78587.880	1139.394	180.000
Std	484751.300	251409.000	4778.828	537.994
Min	7600.000	4000.000	0.000	0.000
25%	18950.000	10075.000	0.000	0.000
50%	37550.000	17400.000	100.000	0.000
75%	78900.000	34500.000	400.000	100.000
Max	3108200.000	1469100.000	29700.000	4000.000

- **Road_55dB**: El ruido de carreteras varía significativamente, con un mínimo de 7,600 y un máximo de 3,108,200, lo que sugiere una amplia variación en la exposición al ruido de carretera entre las diferentes áreas observadas, ademas de acuerdo a la media y la mediana hay indicios de falta de simetria y que hay inclinación a la cola derecha inicialmente algo considerable.
- **Road_60dB**: Similarmente, el de ruido de carreteras varía desde 4,000 hasta 1,469,100, indicando también una gran variabilidad, ademas de acuerdo a la media y la mediana hay indicios de falta de simetria y que hay inclinación a la cola derecha inicialmente fuerte .
- **Railways_65dB**: El ruido de ferrocarriles tiene un mínimo de 0 (indicando áreas sin exposición a este nivel de ruido de ferrocarriles), ademas de acuerdo a la media y la mediana hay indicios de falta de simetria y que hay inclinación a la cola derecha inicialmente algo considerable.
- **Industry_65dB**: Para el ruido industrial a 65 dB varía desde 0 hasta 4,000, con una media de 180, aunque hay un valor faltante en esta columna, ademas de acuerdo a la media y la mediana hay indicios de falta de simetria y que hay inclinación a la cola derecha algo considerable.

3. Realice un breve análisis exploratorio para identificar la distribución de las variables usadas en la base de datos ¿será que existe relación entre las variables?

3.1. Distribución de las variables

Para analizar la distribución de las variables en el dataset(Ruidoso), se emplearán visualizaciones gráficas, específicamente histogramas. Estas representaciones visuales nos permitirán comprender de manera más clara la dispersión y la forma de cada variable, lo que facilitará la identificación de patrones o características importantes en los datos.

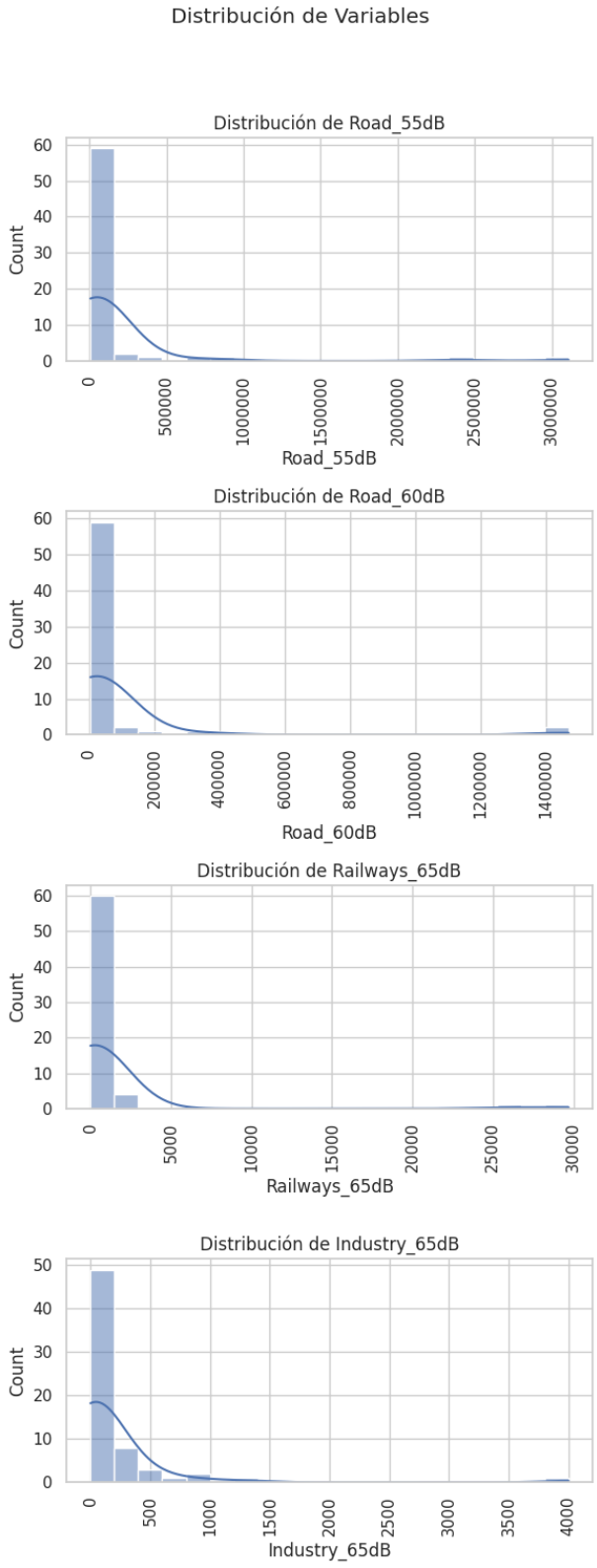


Figure 1. *Histogramas*

- **"Road_55dB"** La mayor concentración de observaciones se sitúa en los valores más bajos, lo que sugiere que, en promedio, las personas están expuestas a niveles bajos de ruido de carretera a 55 dB , ademas de evidenciar que la distribución es muy sesgada hacia la derecha como habiamos analizado antes, con unos pocos valores extremadamente altos.
- **"Road_60dB"**: Se presenta un patrón parecido al de "Road_55dB", con una agrupación de los datos hacia los niveles más bajos y un sesgo hacia la derecha. Esto implica que, aunque la exposición promedio al ruido a 60 dB es baj, pero también con un sesgo hacia la derecha. Nuevamente, hay algunos valores muy altos que podrían ser atípicos.
- **"Railways_65dB":** La exposición al ruido de ferrocarriles a 65 dB se caracteriza por una tendencia similar, con la mayor parte de las mediciones reflejando una exposición baja y pocos casos con exposiciones altas. La inclinación de la distribución hacia la derecha sugiere que, mientras la mayoría de las personas experimentan niveles bajos de ruido de ferrocarriles, existen valores atípicos con exposiciones más altas.

3.2. Caso variable "Industry65dB:"

Como sabemos esta variable contiene un valor faltante, en este caso tenemos 2 soluciones:

- **Eliminación**: Consiste en eliminar la fila del dato faltante.
- **Imputacion**: Es una metodo de relleno de datos ante la presencia de datos faltantes, que podria se a traves de la media, mediana o otra medida vista en la "parte 2" de este trabajo.

Analizaremos cada caso para entender los efectos que podrían tener los datos.

Caso 1: Eliminacion.

Despues de eliminar los NaN's

Table 6. *Eliminación de NaN's*

Columna	NaN's
Road_55dB	0
Road_60dB	0
Railways_65dB	0
Industry_65dB	0

Tanto los resúmenes estadísticos como los histogramas muestran la siguiente forma:

Table 7. *Estadísticas de los datos(Sin NaN's*

	Road_55dB	Road_60dB	Railways_65dB	Industry_65dB
Count	65.000	65.000	65.000	65.000
Mean	113860.000	57195.380	760.000	180.000
Std	317306.200	183079.200	3680.430	537.994
Min	7600.000	4000.000	0.000	0.000
25%	18600.000	10000.000	0.000	0.000
50%	37300.000	17400.000	100.000	0.000
75%	77400.000	33600.000	400.000	100.000
Max	2387200.000	1426100.000	29700.000	4000.000

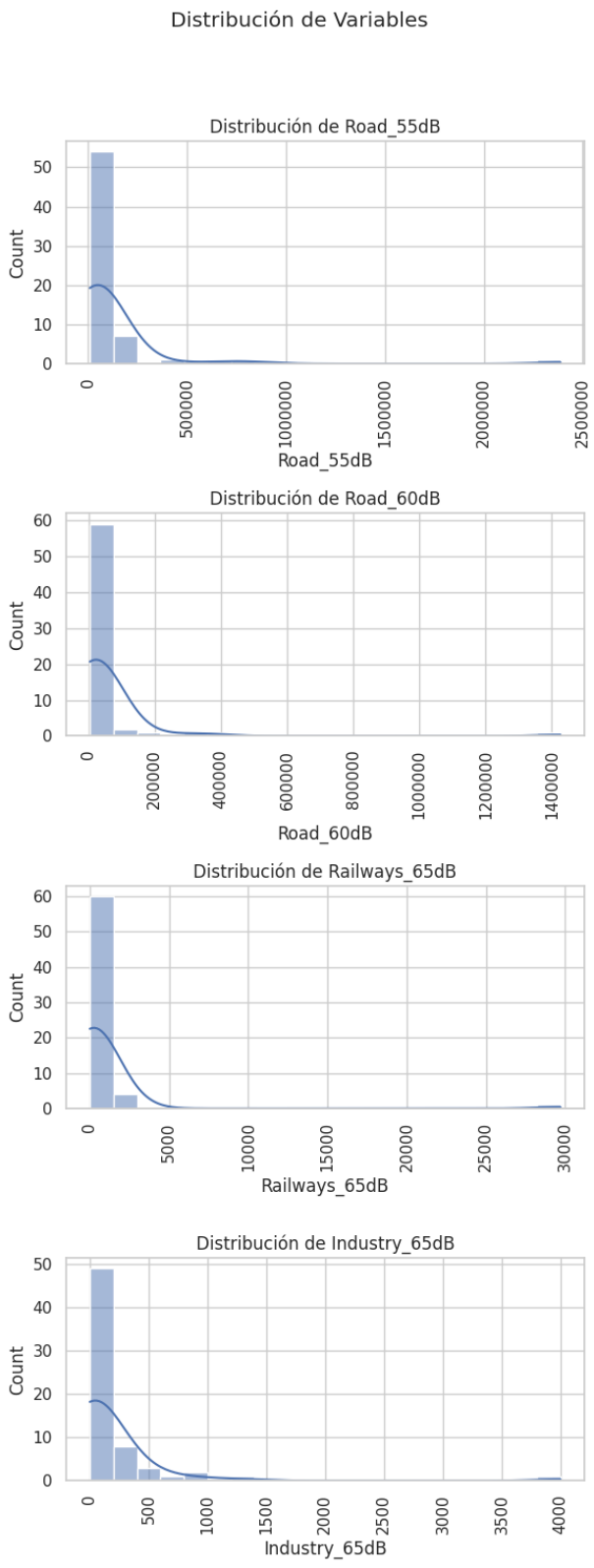


Figure 2. *Histogramas(Sin NaN's)*

Con este método, al comparar los resúmenes estadísticos y los histogramas con los resultados obtenidos con los datos que contienen valores NaN, no se observa ninguna variación significativa. Por lo tanto, la opción de eliminar los valores NaN podría considerarse una alternativa válida y adecuada en este caso.

Caso 2: Imputación

En este caso, se decide no considerar la técnica de imputación de datos, ya que se ha evidenciado que la presencia de datos faltantes no altera los resultados ni modifica la distribución de las variables. Además, al tratarse únicamente de un dato faltante, la imputación se vuelve aún menos relevante.

Conclusión.

Los resultados y conclusiones previos, antes de considerar los casos de eliminación y imputación mediante resúmenes estadísticos y histogramas, siguen siendo válidos. Sin embargo, para evitar posibles errores en las técnicas a utilizar, especialmente aquellas que pueden ser inestables ante la presencia de NaN, se optará por trabajar con el dataset sin valores faltantes a través de la eliminación. Esto garantizará una mayor estabilidad en el análisis y la aplicación de técnicas estadísticas posteriores.

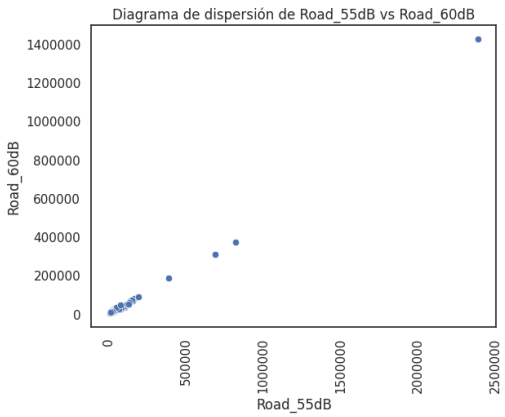
3.3. ¿Cuál es la distribución de los datos?

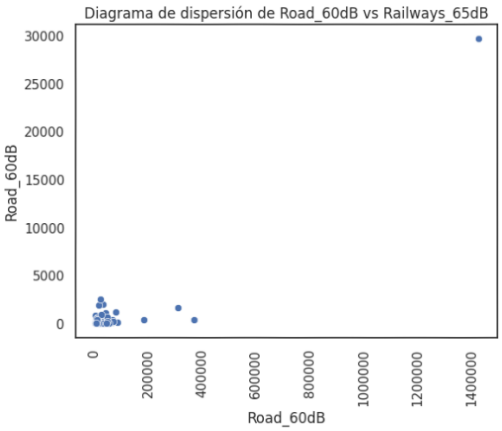
Observando estos histogramas, se concluye que los datos presentan una distribución sesgada hacia la derecha, lo que significa que hay una concentración de valores bajos y una presencia menos frecuente de valores altos. Esto descarta la posibilidad de que sigan una distribución normal, la cual se caracterizaría por una simetría alrededor de una media central. Dada la forma de los datos, podrían ajustarse mejor a una distribución log-normal, exponencial o de Pareto, todas las cuales son capaces de modelar datos con este tipo de asimetría. Estos modelos son útiles especialmente en contextos donde las magnitudes varían ampliamente, como podría ser el caso con la densidad de carreteras, vías férreas e industrias. Por lo tanto, no consideraría una distribución normal para analizar estos datos, sino que buscaría en modelos que manejen de forma efectiva el sesgo a la derecha que he identificado.

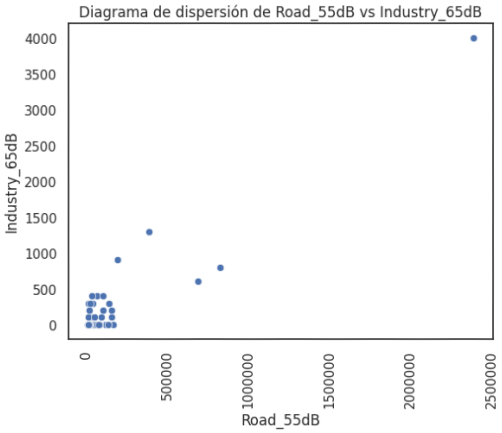
3.4. Relacion de variables

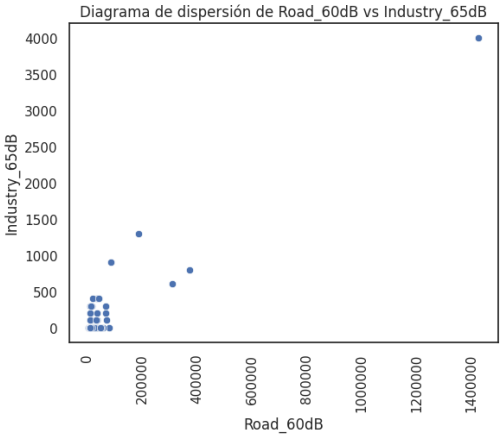
Para evaluar la relación entre las variables, podemos utilizar el coeficiente de correlación de Pearson, que nos proporciona información sobre la correlación lineal entre pares de variables. Esta información puede visualizarse en un mapa de calor. Sin embargo, es importante tener en cuenta que Pearson está diseñado para detectar correlaciones lineales. Por lo tanto, también realizaremos diagramas de dispersión para examinar posibles relaciones no lineales entre las variables.

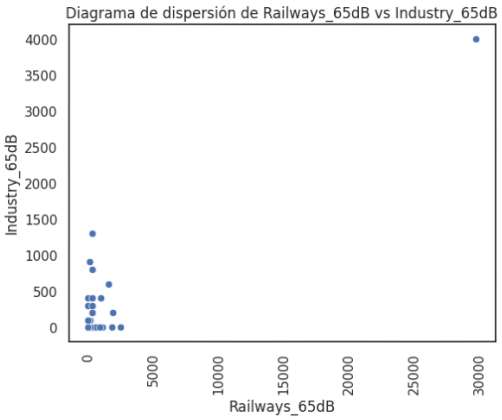
Diagramas de dispersión











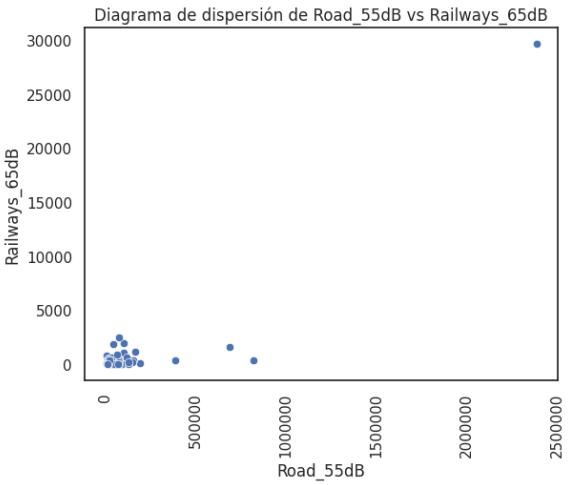


Figure 3. *Diagramas de dispersión*

En los diagramas de dispersión, se observa una correlación lineal solamente entre las variables "Road_55dB" y "Road_60dB". En cuanto a las otras variables, no se aprecia una correlación lineal clara; en su lugar, se observa una concentración de puntos que sugiere una posible relación no lineal o la ausencia de una relación clara entre ellas.

Correlación de pearson:

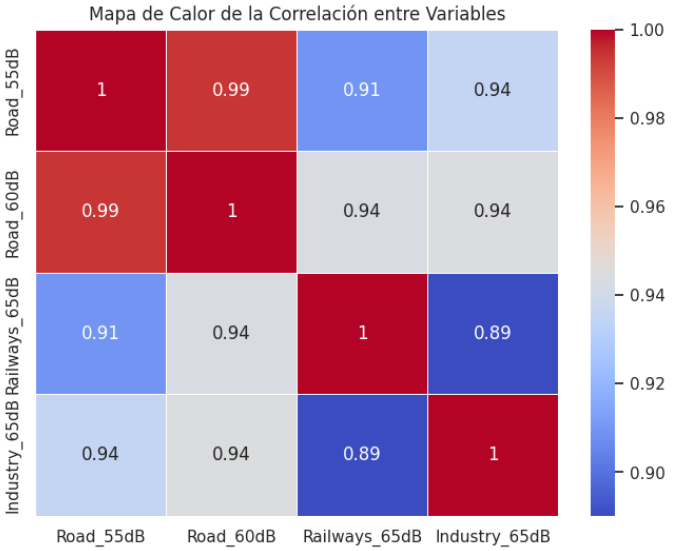


Figure 4. *Correlación de pearson*

Concluiciones indivuales:

- **Road_55dB** y **Road_60dB**: La correlación es extremadamente alta 0.99, lo cual se podria esperar ya que ambos son niveles de ruido de carreteras y es probable que si está expuesta a 55 dB también lo esté a 60 dB.
- **Road_55dB** - **Road_60dB** y **Railways_65dB**: También hay una alta correlación (0.91 y 0.94 respectivamente) entre el ruido de carreteras y ferrocarriles, lo que puede indicar que personas cercanas a carreteras también están frecuentemente cerca de ferrocarriles, o que las fuentes de ruido tienden a agruparse.

- **Road_55dB**- **Road_60dB** y **"Industry_65dB"**: De nuevo, la correlación es muy alta 0.94, sugiriendo una asociación similar a la anterior, donde personas expuestas a ruido de carreteras también lo están a ruido industrial.
- **Railways_65dB** y **Industry_65dB**: Con una correlación de 0.89, es menos fuerte que las demás, pero sigue siendo significativa, lo que sugiere que muchas de las áreas expuestas a ruido de ferrocarriles también están expuestas a ruido industrial.

Otras consideraciones:

- **Agrupación de Fuentes de Ruido**: Es posible que haya áreas que están expuestas a un tipo de ruido tienden a estar expuestas a otros tipos también, lo que podría deberse a la planificación urbana o industrial que concentra varias fuentes de ruido en las mismas zonas.
- **Implicaciones para la Mitigación del Ruido**: Para estrategias de mitigación del ruido, estos resultados sugieren que las intervenciones podrían necesitar considerar múltiples fuentes de ruido simultáneamente, ya que es probable que haya áreas afectadas estén sujetas a una combinación de ruido de carreteras, ferrocarriles e industrias.

Conclusión:

Hay que tener cuidado con las concluiones que se obtuvo de la correlación de person por lo que se evidencio en los digramas de dispersión ya que pearson aunque da una correlacion en todas alta puede que este sesgada ya que como se habia defino pearson se utiliza para relaciones lineales y la relaciones que se observan es mas de agrupación.

4. Verifique si existen problemas de datos atípicos en cada una de las variables usando las metodologías de detección a nivel univariado.

Los valores atípicos pueden manifestarse de manera global, contextual y colectiva en un conjunto de datos. En el caso univariado, la detección de valores atípicos se puede llevar a cabo de las siguientes maneras:

- **Boxplot**: Estos gráficos permiten visualizar valores extremos univariados. **La regla de decisión** comúnmente utilizada es la de +/- 1.5 veces el Rango Intercuartílico (IQR).
- **Z-score**: Esta métrica indica cuántas desviaciones estándar tiene una observación de la muestra, asumiendo una distribución gaussiana. Se calcula según la fórmula:

$$Z_i = \frac{x_i - \mu}{\sigma}$$

  Donde x_i es la observación, μ es la media y σ es la desviación estándar. **La regla de oro** es Z > 3.
- **Z-score Modificado**: La media y la desviación estándar muestrales pueden verse afectadas por los valores extremos presentes en los datos. Por lo tanto, se utiliza el Z-score modificado, calculado de la siguiente manera:

$$M_i = \frac{0.6745(x_i - \tilde{x}_i)}{MAD}$$

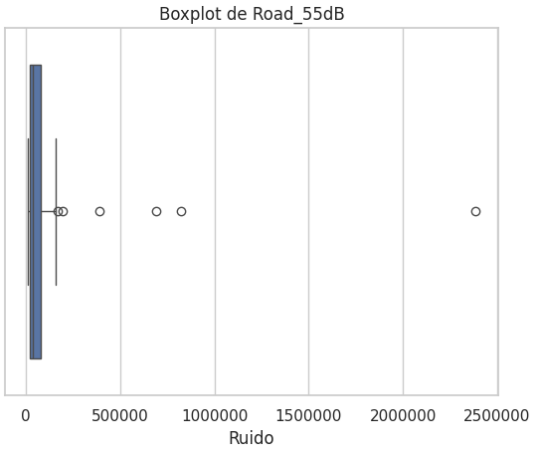
$$MAD = median(|x_i - \tilde{x_i}|)$$

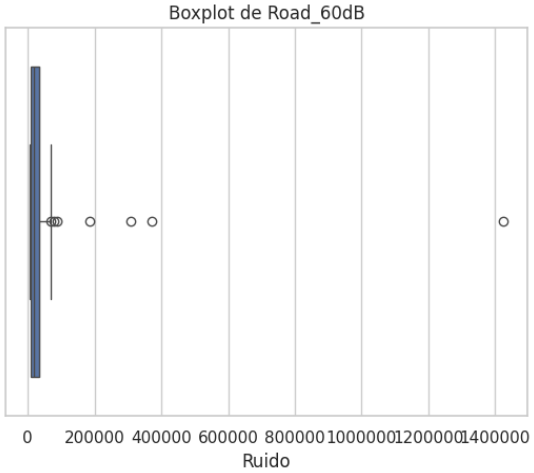
  Donde $\tilde{x}_i$ es la mediana de las diferencias absolutas entre las observaciones y la media, y MAD es la desviación absoluta de la mediana, **La regla de oro** valores mayores 3.5 son datos Outliers

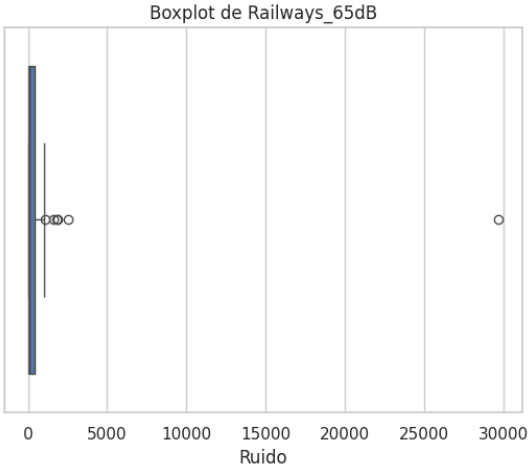
Además de las tres metodologías mencionadas anteriormente, también es posible utilizar histogramas para detectar valores atípicos. En la sección (3.1), aunque se emplearon principalmente para visualizar la forma de la distribución, los histogramas también nos ofrecen la oportunidad de identificar posibles valores atípicos. A primera vista, de acuerdo con los histogramas, parece haber presencia de datos atípicos. Por lo tanto, los tres métodos mencionados anteriormente nos serán útiles para establecer una regla de decisión con el fin de determinar qué valores se consideran como outliers.

Observación: Recordemos que los datos atípicos pueden surgir por una variedad de razones, y no necesariamente son errores. Por lo tanto, al detectar outliers, es importante investigar la razón detrás de estos datos. Lamentablemente, en este caso, puede ser un poco complejo determinar la razón, ya que, como se menciona en la sección (2.1), no tenemos certeza sobre la escala de medida utilizada.

4.1. Boxplot







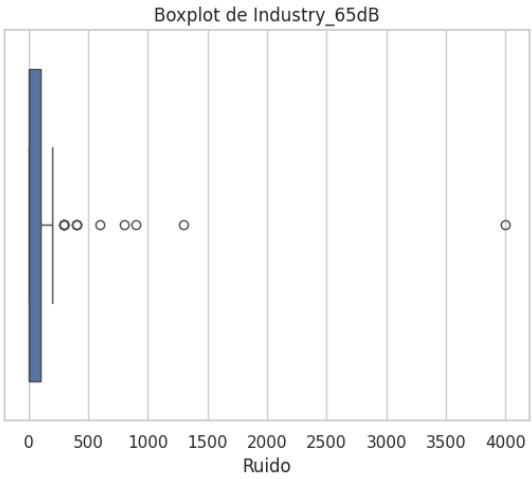


Figure 5. *Boxplots*

4.2. Z-score

Table 8. *Outliers Road_55dB*

Fila	Road_55dB	Road_60dB	Railways_65dB	Industry_65dB
48	2387200	1426100	29700	4000.00
Total:1				

Table 9. *Outliers Road_60dB*

Fila	Road_55dB	Road_60dB	Railways_65dB	Industry_65dB
48	2387200	1426100	29700	4000.00
Total:1				

Table 10. *Outliers Railways_65dB*

Fila	Road_55dB	Road_60dB	Railways_65dB	Industry_65dB
48	2387200	1426100	29700	4000.00
Total:1				

Table 11. *Outliers Industry_65dB*

Fila	Road_55dB	Road_60dB	Railways_65dB	Industry_65dB
48	2387200	1426100	29700	4000.00
Total:1				

4.3. Z-score modificado

Table 12. *Outliers Road_55dB*

Fila	Road_55dB	Road_60dB	Railways_65dB	Industry_65dB
0	166400	79200	1100	0.00
7	388000	185200	300	1300.00
12	192900	86000	100	900.00
14	823200	371700	300	800.00
26	156000	67300	300	200.00
28	689300	309300	1600	600.00
48	2387200	1426100	29700	4000.00
Total: 7				

Table 13. *Outliers Road_60dB*

Fila	Road_55dB	Road_60dB	Railways_65dB	Industry_65dB
0	166400	79200	1100	0.00
7	388000	185200	300	1300.00
12	192900	86000	100	900.00
14	823200	371700	300	800.00
17	140900	66200	100	300.00
26	156000	67300	300	200.00
28	689300	309300	1600	600.00
44	152500	69700	200	100.00
48	2387200	1426100	29700	4000.00)
Total: 9				

Table 14. *Outliers Railways_65dB*

Fila	Road_55dB	Road_60dB	Railways_65dB	Industry_65dB
0	166400	79200	1100	0.00
4	103100	33500	1900	200.00
10	57900	23300	700	0.00
28	689300	309300	1600	600.00
29	7600	4000	800	0.00
31	102800	41600	1000	400.00
40	47500	17400	1800	0.00
48	2387200	1426100	29700	4000.00
49	77200	24300	2500	0.00
55	64300	25100	900	0.00)
Total: 10				

Table 15. *Outliers Industry_65dB*

Fila	Road_55dB	Road_60dB	Railways_65dB	Industry_65dB
3	24500	16700	0	100.00
4	103100	33500	1900	200.00
5	36800	16100	0	100.00
6	39700	16800	0	300.00
7	388000	185200	300	1300.00
8	15000	9400	300	200.00
9	18600	12000	0	100.00
11	92700	33600	0	100.00
12	192900	86000	100	900.00
14	823200	371700	300	800.00
17	140900	66200	100	300.00
18	38000	13400	100	300.00
19	14100	10000	0	300.00
20	64800	36400	0	400.00
25	32400	19700	300	400.00
26	156000	67300	300	200.00
28	689300	309300	1600	600.00
31	102800	41600	1000	400.00
37	31600	10700	200	100.00
44	152500	69700	200	100.00
48	2387200	1426100	29700	4000.00
53	11600	8300	100	100.00
57	49700	32900	0	100.00
63	24500	12400	300	300.00

Total: 24

4.4. Conclusión

Individuales

- **Boxplot**: En los gráficos de la sección (4.1), se observa una forma muy similar con la presencia de varios datos atípicos. Además, en todos los boxplot se destaca la presencia de un dato notablemente distante del resto.
- **Z-score**: En las tablas de la Subsección (4.2), al aplicar esta medida para detectar valores atípicos, se encontró únicamente un dato fuera de lo común. Es importante destacar que este valor atípico coincide con el que se identificó como el más alejado mediante el método de los boxplot, y lo notable es que este mismo dato atípico se repite en todas las variables analizadas.
- **Z-score modificado**:El método discutido en la Subsección (4.3) identificó una mayor cantidad de valores atípicos, algunos de los cuales eran consistentes en todas las variables analizadas. Es relevante resaltar que la variable Industry_65dB mostró una detección más pronunciada de datos atípicos, lo cual concuerda con la observación de los boxplot, donde se observaron más outliers en todas las variables.

Grupales

- Los **boxplot** y el **Z-score** coinciden porque ambos métodos están fundamentados en la premisa de que los datos se distribuyen de manera simétrica.
- El **Z-score** y el **Z-score modificado** difieren considerablemente debido a su aplicabilidad en diferentes condiciones de distribución de los datos. Mientras que el z-score se aplica en situaciones donde los datos muestran simetría, el z-score modificado se utiliza cuando existe un sesgo hacia una cola en los datos, lo que resulta en una discrepancia entre la media y la mediana.

Nota: Se consideró el método del Rango Intercuartil (**IQR**) como otra medida para detectar outliers; sin embargo, sus resultados fueron muy similares a los outliers mostrados en los boxplot, por lo que no se presentaron en el análisis final.

4.5. ¿Cuál método resulta más eficaz?

Tras analizar las conclusiones en la sección (5.4), se evidencia que el método más efectivo para identificar outliers es el Z-score modificado. Este método sobresale por su capacidad para detectar un mayor número de valores atípicos en comparación con los otros métodos evaluados. Además, se sugiere que el Z-score modificado habría sido la elección óptima desde el principio, ya que los histogramas presentados en la sección (3.1) revelaban una distribución no simétrica de los datos.

5. ¿Se detectan valores atípicos a nivel multivariado?

En el análisis multivariado, la detección de outliers puede realizarse utilizando diversas metodologías, que incluyen:

- **Distancia de Manhalobis**: Es una medida de distancia entre puntos $\vec{x} = (x_1, x_2, x_3)$ y el vector de medias $\vec{\mu} = (\mu_1, \mu_2, \mu_3)$, teniendo en cuenta la matriz de varianzas y covarianzas **S**. La distancia de Manhalobis se define como:

$$D_m(\vec{x}) = \sqrt{(\vec{x} - \vec{\mu})^\tau S^{-1} (\vec{x} - \vec{\mu})}$$

- **Local Outlier Factor(LOF)**:Local outlier factor es un método basado en **densidad** que utiliza la búsqueda de vecinos más cercanos. Ese método calcula los **scores** para cada uno de los puntos a partir de la tasa promedio de densidad de los puntos vecinos con respecto a si mismo.

$$lrd_k(o) = \frac{||N_k(0)}{\sum_{0' \epsilon N_k(o)} reachdistk(o' \leftarrow o)}$$

$$LOF_k(o) = \sum_{0' \epsilon N_k(o)} lrd_k(o') * \sum_{0' \epsilon N_k(o)} readchdist_k(o' \leftarrow o)$$

  - $N_k(o) = \{o' \mid o' \ en \ D, \ dist(o, o') \leq dist_k(o)\}$
  - reachdistk$(o' \leftarrow o) = max(dist_k(o), dist(o, o'))$
  - $dist_k(o)$= k-Distancia, es una distancia entre o y **k**-ésimo vecino mas cercano (kNN)

- **Isolation Forest**: Las instancias de datos anómalas se pueden aislar de los datos normales mediante la partición recursiva del conjunto de datos. **Score**

$$2^{\frac{-E(h_{(x)})}{c(n)}}$$

  Donde:

  - $h_{(x)}$: Es la longitud del camino a x.
  - $E(h_{(x)})$: Es la media de las alturas de x en todos los iForest.
  - c(n):es una costante de normalización puede ser calculado como:

$$c(n) = 2H(n-1) - (2(n-1)/n)$$

  - H(i): es el numero armónico:

$$H(i) = ln(i) + 0.5772156649(constante\ euler)$$

5.1. Distancia Mahalanobis

Table 16. *Mahalanobis*

Fila	Road_55dB	Road_60dB	Railways_65dB	Industry_65dB	Mahalanobis
48	2387200	1426100	29700	4000.00	7.92
14	823200	371700	300	800.00	5.80
7	388000	185200	300	1300.00	4.97
28	689300	309300	1600	600.00	4.53
49	77200	24300	2500	0.00	4.08
12	192900	86000	100	900.00	3.82
4	103100	33500	1900	200.00	3.65
40	47500	17400	1800	0.00	2.56
31	102800	41600	1000	400.00	2.27
60	74400	47000	0	0.00	2.12

Las 10 distancias más grandes.

5.2. Local Outlier Factor(LOF)

Table 17. Lof_score

Fila	Road_55dB	Road_60dB	Railways_65dB	Industry_65dB	Lof_score
48	2387200	1426100	29700	4000.00	37.89
14	823200	371700	300	800.00	12.33
28	689300	309300	1600	600.00	10.21
7	388000	185200	300	1300.00	5.87
12	192900	86000	100	900.00	2.57
0	166400	79200	1100	0.00	2.21
26	156000	67300	300	200.00	2.02
44	152500	69700	200	100.00	2.00
17	140900	66200	100	300.00	1.87
58	132400	56100	0	0.00	1.75

Las 10 distancias más grandes.

5.3. Isolation Forest

Table 18. $puntajes_anomalia$

Fila	Road_55dB	Road_60dB	Railways_65dB	Industry_65dB	Puntaje
48	2387200	1426100	29700	4000.00	0.89
28	689300	309300	1600	600.00	0.70
14	823200	371700	300	800.00	0.70
7	388000	185200	300	1300.00	0.67
12	192900	86000	100	900.00	0.59
0	166400	79200	1100	0.00	0.56
4	103100	33500	1900	200.00	0.55
49	77200	24300	2500	0.00	0.54
31	102800	41600	1000	400.00	0.51
26	156000	67300	300	200.00	0.48

Las 10 distancias más grandes(Con 100 árboles).

5.4. Conclusión

A pesar de las diferencias en los métodos y las métricas empleadas, se observó una coincidencia notable: los mismos cuatro datos ocuparon las cuatro primeras posiciones en la detección de outliers, aunque en diferentes posiciones para cada método. Esto sugiere que estos cuatro datos podrían ser considerados outliers con mayor seguridad en un análisis multivariado. La consistencia en obtener las puntuaciones más altas en sus respectivos métodos, en comparación con otros datos, refuerza su importancia y resalta su singularidad dentro del conjunto de datos.

6. Para el caso univariado, escoja una variable y realice un análisis sobre lasimplicaciones que tiene realizar diferentes tratamientos a los datos atípicos en la distribución de la respectiva variable.

Como se observó en la sección (4), donde se analizaron los datos atípicos univariados utilizando el método del **Z-score modificado**, se encontró que la variable 'Industry_65dB' presentó el mayor número de casos de outliers. Debido a esta observación, se enfocará el análisis posterior en esta variable específica.

Variable a analizar: Industry_65dB

Después de identificar los outliers, es importante considerar cómo gestionarlos. Vamos a analizar cuatro métodos para abordarlos:

- **Mantenerlos**:Si consideramos que los outliers pueden ser representativos de un subconjunto significativo de nuestros datos, podemos optar por mantenerlos en el análisis.
- **Eliminarlos**: Si tenemos certeza de que los outliers son resultado de un error en la entrada de los datos, como un error humano o de medición, y no podemos corregirlo, entonces podemos optar por eliminarlos del conjunto de datos.
- **Imputarlos**: La imputación implica sustituir los valores atípicos por otros valores, como la mediana o la media. Esto se realiza generalmente cuando deseamos conservar la mayor cantidad de datos posible, al mismo tiempo que eliminamos el efecto de los outliers.
- **Winzorizar**:La winsorización es una técnica que sustituye los valores atípicos por el valor más cercano que no se considera un outlier según ciertos criterios.

6.1. Metodos

6.1.1. Mantenerlos

Este método no se considera una opción viable para el contexto del dataset, ya que se entiende que estos datos son extremadamente atípicos en el contexto de la medición del ruido. Sin embargo, es importante tener en cuenta que la medida de los datos puede tener cierto grado de incertidumbre en cuanto a la escala de medida utilizada.

6.1.2. Eliminarlos

Con este metodo al hacer un resumen numerico podemos ver como cambio la variable:

Resumen estadistico

Table 19. *Comparaciém*

	Antes	Despues
Count	65.000	58.00
Mean	180.000	67.24
Std	537.994	120.51
Min	0.000	0.00
25%	0.000	0.00
50	0.000	0.00
75%	100.000	100.00
Max	4000.000	400.00

Histogramas

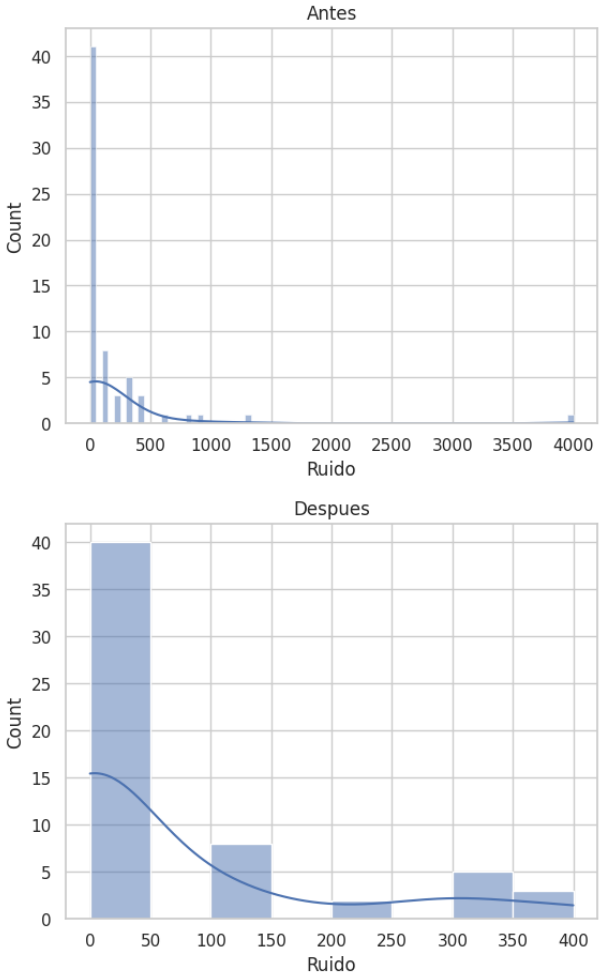


Figure 6. *Comparación*

- **Antes**: La varible Industry_65dB sin elimanción de Outliers
- **Despues**:La variable Industry_65dB sin Outliers

Diagrama de cajas

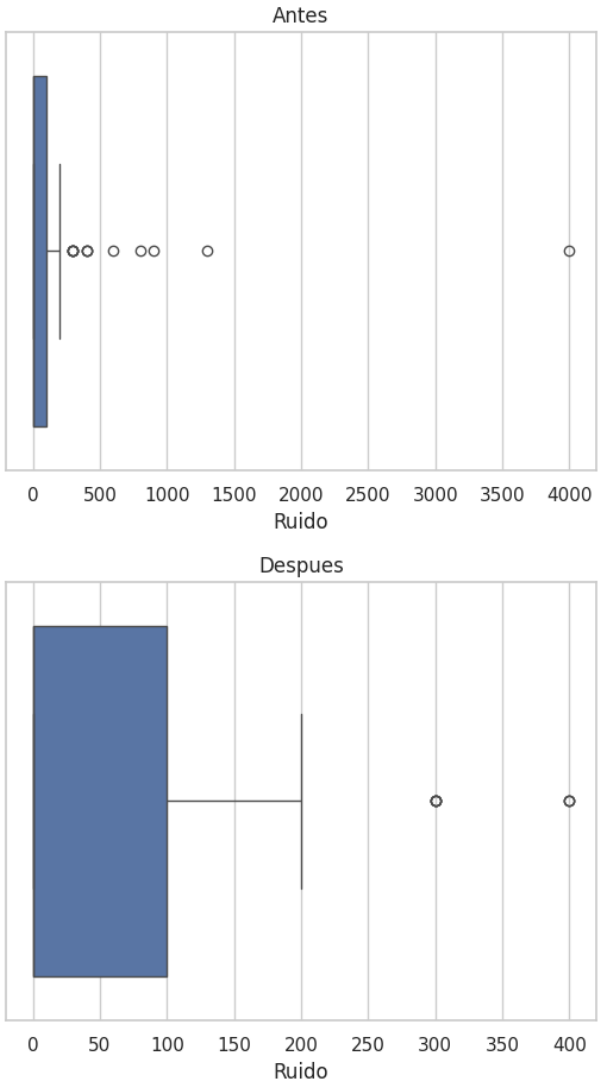


Figure 7. *Comparación*

- **Antes**: La variable Industry_65dB sin elimanción de Outliers
- **Despues**:La variable Industry_65dB sin Outliers

La comparación entre los dos histogramas muestra un cambio drástico en la distribución del ruido antes y después de eliminar los datos atípicos. En el histograma "Antes", observamos una distribución con una cola muy larga hacia la derecha, lo que indica la presencia de valores extremos que extienden la escala del eje x hasta cerca de 4000.

En el histograma "Después", una vez que se eliminan los datos atípicos, la escala del eje x es mucho más pequeña, con un límite cercano a 400, reflejando una distribución mucho más concentrada. La mayoría de las observaciones están agrupadas cerca del extremo inferior de la escala, lo que indica que la exposición al ruido es generalmente baja para la mayoría de las observaciones sin los valores extremos.

Al eliminar los datos atípicos, las métricas estadísticas como la media y la mediana se vuelven más cercanas; sin embargo, en el resumen estadístico se observa que aún están ligeramente distanciadas.

También se observa que en los boxplot hubo una reducción en la cantidad de datos atípicos después de eliminarlos del conjunto de datos.

6.1.3. Imputación

En este caso, como se evidencia en el resumen estadístico (2.4), los datos están sesgados hacia la cola derecha. Por lo tanto, para la imputación de valores atípicos, se utilizará la mediana.

Resumenes estadisticos

Table 20. *Comparaciém*

	Antes	Despues
Count	65.000	65.00
Mean	180.000	53.85
Std	537.994	111.91
Min	0.000	0.00
25%	0.000	0.00
50	0.000	0.00
75%	100.000	0.00
Max	4000.000	400.00

Histogramas

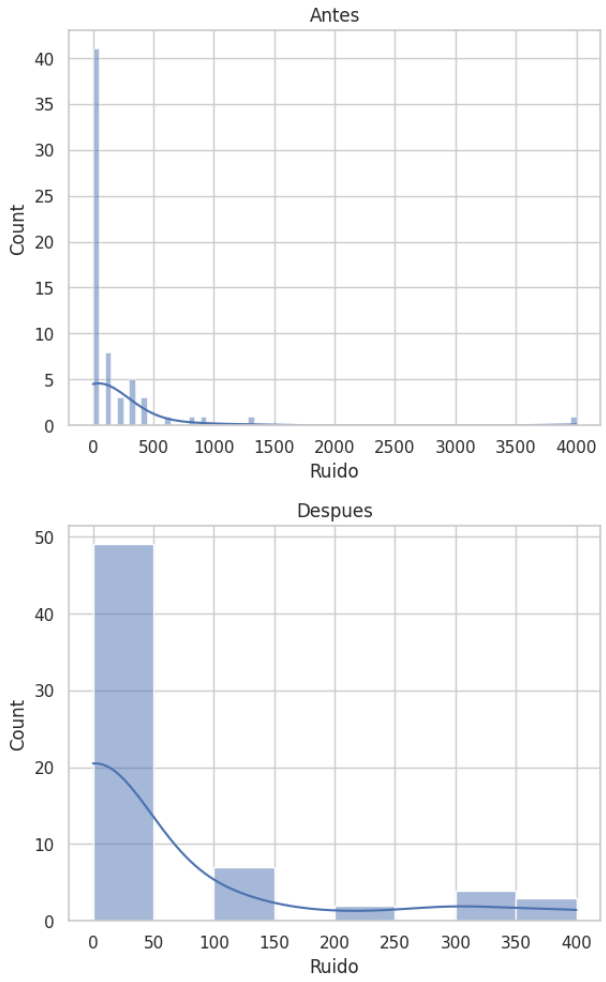


Figure 8. *Comparación*

Diagramas de cajas

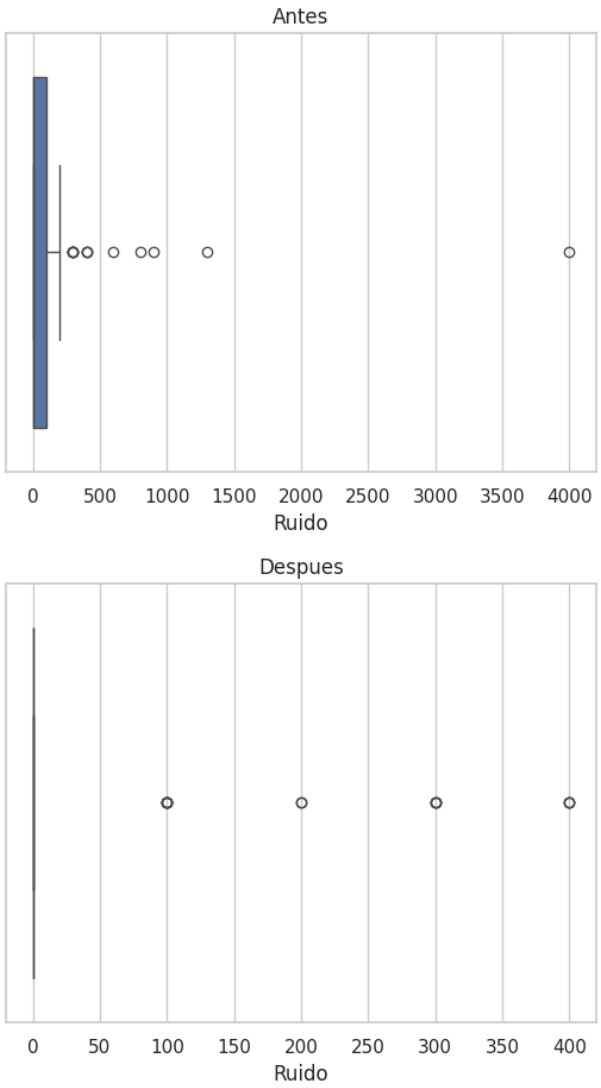


Figure 9. *Comparación*

La técnica de imputación de la mediana aplicada al conjunto de datos ha transformado sustancialmente la distribución y las estadísticas resumidas del ruido registrado. Antes de la intervención, los histogramas y los boxplots revelaban una dispersión considerable, caracterizada por una amplia gama de valores y la presencia prominente de datos atípicos que se extendían hasta 4000. Estos valores extremos, al ser tan distantes de la mayoría de las observaciones, ejercían una influencia desproporcionada sobre la media y la desviación estándar, inflando estas medidas y, por tanto, distorsionando la interpretación real del nivel de ruido al que están expuestas la mayoría de las personas.

Tras la imputación, los histogramas muestran una distribución más estrecha y centrada, con una eliminación efectiva de la cola larga que indicaba la presencia de valores extremos. Correspondientemente, los boxplots actualizados reflejan una distribución mucho más compacta,la imputación de la mediana ha reducido significativamente la media de aproximadamente 180 a 53.85 y la desviación estándar de 537.99 a 111.91, lo cual indica una consistencia y centralización notables en los niveles de ruido. Las estadísticas resumidas ahora muestran una mediana y percentiles que reflejan una agrupación más cercana al valor de ruido más frecuente, y el máximo ha disminuido drásticamente de 4000 a 400, lo que confirma que la mayoría de los datos se concentran en un intervalo de ruido mucho menor.

6.1.4. Winzorizar

Para esta ultimo metodo se establece los límites de winsorización para el 10% de los valores en ambos extremos de la distribución. Es decir, el 10% de los valores más bajos y el 10% de los valores más altos serán reemplazados por los valores en el percentil 10 y el percentil 90, respectivamente, **Observacion:** se tomo la desición de un limite del 10% por que la cola derecha era muy larga, ademas cuanto se intento con 0.05% los resultados no fueron favorable como los demas metodos, por lo cual se decidio un limite del 0.1% con esto los resultados fueron mu parecido a cuando se utilizo el metodo de elimanacion por lo que solo se analizara el boxplot

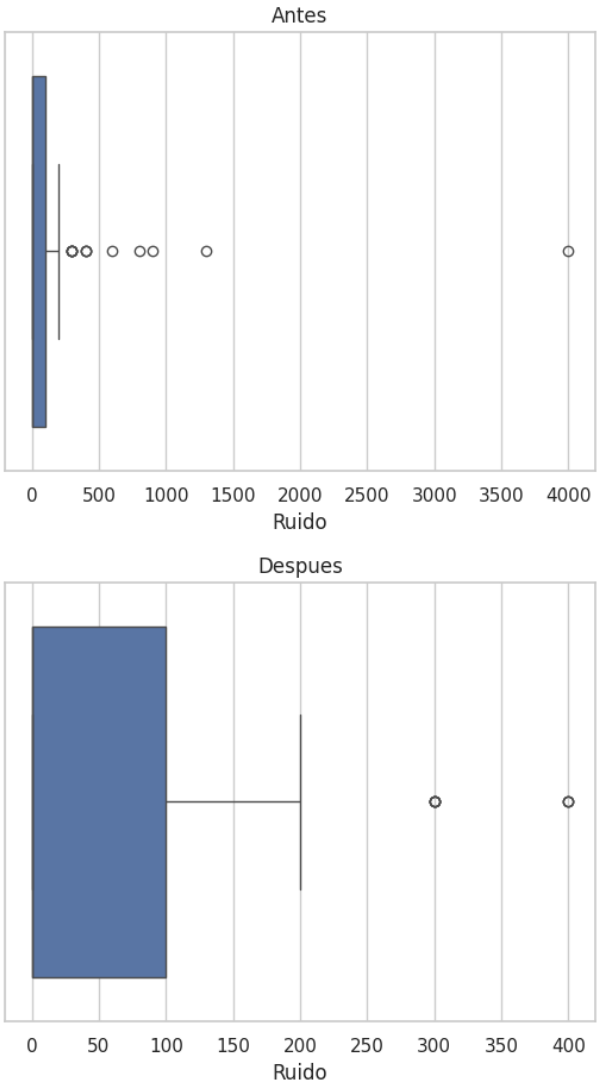


Figure 10. *Comparación*

Se aprecia una reducción en los datos atípicos, y las conclusiones son altamente congruentes con las obtenidas mediante el método de eliminación. Por ende, no es necesario extenderse más en la explicación, ya que serían esencialmente las mismas que las presentadas en la sección (6.1.1).

6.2. conclusión

Los efectos diversos que pudo haber experimentado la variable Industry_65dB se pudieron observar mediante varios métodos. Tanto la eliminación como la winsorización mostraron resultados muy similares, lo cual puede explicarse debido a que la winsorización, en esencia, implica la eliminación de datos. Además, el método de imputación no solo redujo los valores atípicos, sino que también concentró los datos alrededor de cero, dado que la imputación se basó en la media, la cual era cero.

7. Conclusión parte A

Después de realizar varios análisis de los métodos aplicados en el contexto de la base de datos sobre el ruido percibido por personas en diversos entornos como industrias, carreteras o ferrocarriles, se puede observar que la mayoría de las variables están concentradas alrededor de cero, lo que indica que o bien el ruido no era muy intenso o que las personas no percibían ruido. Sin embargo, los datos mostraron una distribución sesgada hacia la cola derecha, lo que sugiere que había personas que percibían una intensidad considerable de ruido u otra medida, dado que la escala de medición está en incertidumbre pero se sabe que trata sobre el ruido. Estos valores tan altos eran indicios de datos atípicos, lo cual llevó a realizar diagnósticos tanto univariados como multivariados para identificar outliers, revelando una gran cantidad de ellos.

Además, se llevó a cabo un análisis sobre cómo afectarían los métodos de tratamiento de outliers en una variable específica, lo que permitió reducir significativamente el rango de ruido. A partir de este resultado, podemos inferir que aplicando este mismo método a las demás variables se podría llegar a resultados similares.

7.1. ¿Cuál cree que es el mecanismo inherente a esos datos atipicos?

Aunque no se aborda en ninguna de las preguntas anteriores, y además, este tema no corresponde a la Parte 1 sino a la Parte 2, que se centra en los datos faltantes, resulta crucial comprender el mecanismo subyacente de los datos atípicos. Estos pueden clasificarse en tres categorías previamente mencionadas: global, contextual y colectiva. Para identificar el mecanismo subyacente, podemos referirnos a la sección 3.4 donde se realizaron diagramas de dispersión. A través de estos, es posible definir las variables que se ajustan a las tres categorías mencionadas de la siguiente manera:

- **Global**: Incluye "Road_55dB" y "Road_60dB".
- **Contextual**: En esta categoría, no se identificaron variables.
- **Colectivo**: Se observaron patrones atípicos en las siguientes combinaciones:
  - "Road_60dB/Railways_65dB".
  - "Road_60dB/Railways_65dB".
  - "Road_60dB/industry_65dB".
  - "Railways_65dB/industry_65dB".
  - "Road_55dB/industry_65dB".

8. Recomendaciones

El principal desafío que enfrentan todos los resultados anteriores radica en sus conclusiones. Aunque las técnicas estadísticas aplicadas están correctas, es importante tener en cuenta que la escala de medición, al estar en incertidumbre, no concuerda completamente con la definición de medición del ruido explicada en la sección (1.1). Esto significa que las conclusiones podrían ser incorrectas al interpretar mal la escala de medida. Por lo tanto, se recomienda publicar la fuente de la base de datos, ya que esto proporcionaría más material para investigar sobre la escala de medida. Además, sería beneficioso comunicarse con el propietario del conjunto de datos para obtener una definición clara de la escala de medida y así evitar posibles interpretaciones erróneas.

Parte 2

9. Sobre el dataset dataset auto-mpg.data-original.txt.

El conjunto de datos 'Auto MPG' contiene una compilación de especificaciones y medidas de eficiencia de combustible para diferentes de automóviles desde finales de la década de 1970 hasta principios de la de 1980. Con atributos que van desde el peso y la potencia hasta el origen y el año del modelo, este conjunto de datos ha sido un recurso para analistas y científicos de datos en la exploración de relaciones entre las características de los vehículos y su eficiencia de combustible. Utilizado ampliamente en el aprendizaje automático, este conjunto de datos no solo ayuda a predecir el consumo de combustible con base en las especificaciones del vehículo, sino que también ofrece perspectivas para la optimización de diseños automotrices hacia una mayor sostenibilidad y eficiencia energética.

10. Cargue y explore el dataset explicando en qué consiste y las características que posee el mismo.

Dataset

Table 21. *auto-mpg.data-original*

MPG	Cylinders	Displacement	Horsepower	Weight	Acceleration	Model Year	Origin	Car Name
18.0	8.0	307.0	130.0	3504.0	12.0	70.0	1.0	chevrolet chevelle malibu
15.0	8.0	350.0	165.0	3693.0	11.5	70.0	1.0	buick skylark 320
18.0	8.0	318.0	150.0	3436.0	11.0	70.0	1.0	plymouth satellite
16.0	8.0	304.0	150.0	3433.0	12.0	70.0	1.0	amc rebel sst
17.0	8.0	302.0	140.0	3449.0	10.5	70.0	1.0	ford torino
27.0	4.0	140.0	86.0	2790.0	15.6	82.0	1.0	ford mustang gl
44.0	4.0	97.0	52.0	2130.0	24.6	82.0	2.0	vw pickup
32.0	4.0	135.0	84.0	2295.0	11.6	82.0	1.0	dodge rampage
28.0	4.0	120.0	79.0	2625.0	18.6	82.0	1.0	ford ranger
31.0	4.0	119.0	82.0	2720.0	19.4	82.0	1.0	chevy s-10

Figure 11. *auto-mpg.data-original*

Este conjunto de datos representa una versión adaptada del original disponible en la biblioteca StatLib. Siguiendo el enfoque de Ross Quinlan en 1993 para la predicción del consumo de combustible (medido en millas por galón), se han excluido 8 casos del conjunto de datos original debido a que presentaban valores desconocidos en la variable dependiente "mpg". Los datos íntegros pueden encontrarse en el archivo "auto-mpg.data-original". La colección de datos se enfoca en el consumo de combustible en condiciones urbanas (mpg) y puede ser estimado mediante 3 atributos discretos multivalor y 5 atributos continuos, según la investigación de Quinlan en 1993.

Las variables que representan el dataset son:

- **MPG(Millas por galón)**: Representa la eficiencia de combustible del vehículo.
- **Cylinders**: Número de cilindros del motor del vehículo
- **Displacement**: Volumen de desplazamientos del motor, generalmente en pulgadas.
- **Horsepower**: Caballos de fuerza del motor.
- **Weight**: Peso del vehículo en libras
- **Acceleration**: Tiempo en segundos que el vehiculo tarda en acelerar de 0 a 60
- **Model Year**: Año de fabricación del vehículo.
- **Origin**: Representa la región de origen del vehículo.
- **Car Name**: Nombre o modelo del automovil.

10.1. ¿cuantos registros hay ?

Table 21. *Registros*

Columna	Cantidad no nula
MPG	398
Cylinders	406
Displacement	406
Horsepower	400
Weight	406
Acceleration	406
Model Year	406
Origin	406
Car Name	406
Total	3654

El conjunto de datos consta de 3654 registros, sin embargo, no todas las columnas cuentan con 406 registros. En este caso, aún no se procederá a revisar la cantidad de valores NaN, como se hizo en la "parte 1", ya que este proceso se abordará en una sección separada específica para ello.

10.2. Tipo de datos

Table 22. *Tipo*

Columna	Type
MPG	float
Cylinders	float
Displacement	float
Horsepower	float
Weight	float
Acceleration	float
Model Year	float
Origin	float
Car Name	object

10.2.1. Limpieza de datos.

En esta etapa de limpieza del conjunto de datos, es necesario corregir el tipo de datos, ya que la forma en que Python lo está leyendo contradice la definición de variables según la fuente del conjunto de datos. Para establecer una referencia clara sobre cómo están definidas las variables, podemos consultar el siguiente enlace: https://archive.ics.uci.edu/dataset/9/auto+mpg, donde se especifica la definición de las variables.

- **displacement**: Continuo(float)
- **mpg**: Continuo(float)
- **cylinders**: Entero(int)
- **horsepower**: Continuo(float)
- **weight**: Continuo(float)
- **acceleration**: Continuo(float)
- **model_year**: Entero(int)
- **origin**: Entero(int)
- **car_name**: Categorica (object)

Table 23. *Tipo corregido*

Columna	Type
MPG	float
Cylinders	int
Displacement	float
Horsepower	float
Weight	float
Acceleration	float
Model Year	int
Origin	int
Car Name	category

Con estos ajustes, ahora es posible realizar análisis de datos, dado que se han corregido los formatos de los mismos.

10.3. Analisis inicial

	MPG	Cylinders	Displacement	Horsepower
count	398.000000	406.000000	406.000000	400.000000
mean	23.514573	5.475369	194.779557	105.082500
std	7.815984	1.712160	104.922458	38.768779
min	9.000000	3.000000	68.000000	46.000000
25%	17.500000	4.000000	105.000000	75.750000
50%	23.000000	4.000000	151.000000	95.000000
75%	29.000000	8.000000	302.000000	130.000000
max	46.600000	8.000000	455.000000	230.000000

Table 24. *Resumen estadistico. parte 1*

	Weight	Acceleration	Model Year	Origin
count	406.000000	406.000000	406.000000	406.000000
mean	2979.413793	15.519704	75.921182	1.568966
std	847.004328	2.803359	3.748737	0.797479
min	1613.000000	8.000000	70.000000	1.000000
25%	2226.500000	13.700000	73.000000	1.000000
50%	2822.500000	15.500000	76.000000	1.000000
75%	3618.250000	17.175000	79.000000	2.000000
max	5140.000000	24.800000	82.000000	3.000000

Table 25. *Resumen estadistico. parte 2*

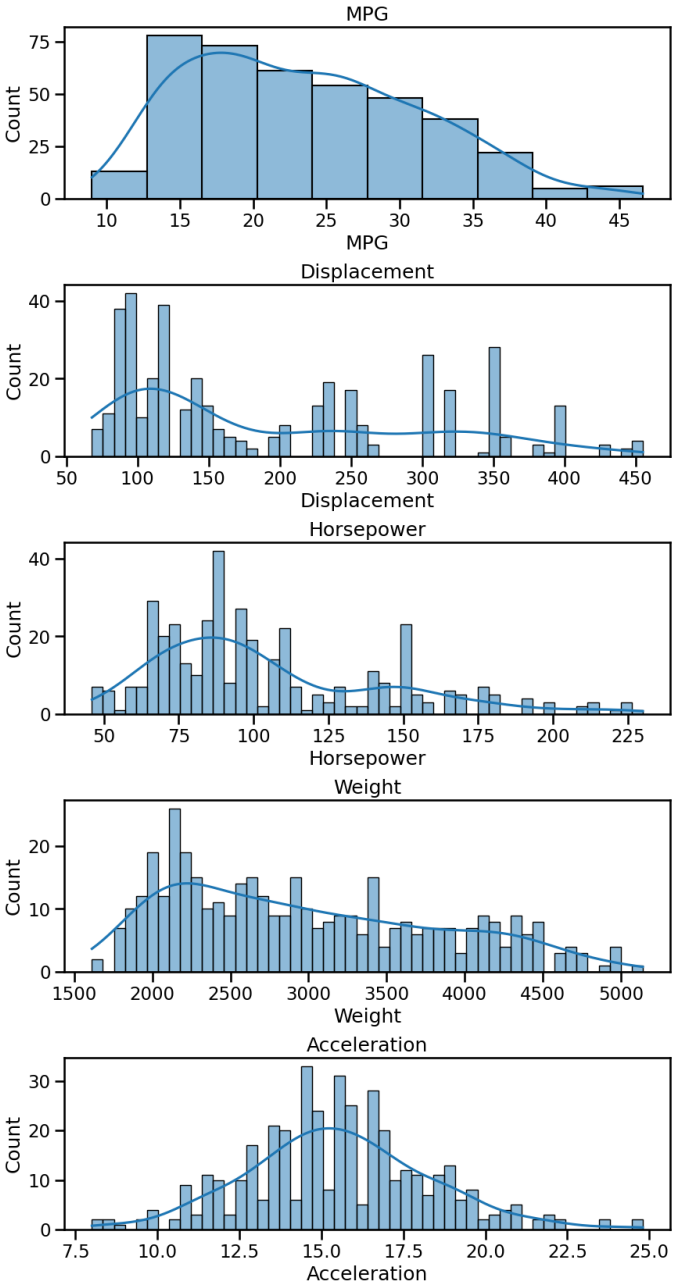
- **MPG** : El rango de MPG va desde 9 hasta 46.6, con una media de 23.51, lo que indica una variabilidad significativa en la eficiencia del combustible entre los vehículos. La desviación estándar de 7.82 sugiere una dispersión moderada en los valores de MPG.
- **Cylinders**: La mayoría de los vehículos tienen entre 3 y 8 cilindros, con un promedio cercano a 5.48, reflejando una distribución bimodal probable debido a la presencia de vehículos con eficiencia de combustible y aquellos enfocados en el rendimiento.
- **Displacement**: El desplazamiento varía ampliamente de 68 a 455 cc, con una media de 194.78 cc. Esto indica una gama amplia de tamaños de motor en el conjunto de datos, desde motores pequeños y eficientes hasta grandes motores de alto rendimiento.
- **Horsepower**: Hay una variedad de potencia en los vehículos, desde 46 hasta 230 caballos de fuerza, con un promedio de 105.08. La distribución sugiere una inclinación hacia vehículos con potencia moderada.
- **Weight**:El peso de los vehículos varía de 1613 a 5140 libras, con una media de 2979.41 libras, lo que indica una diversidad de tamaños y tipos de carrocería, desde compactos ligeros hasta vehículos más grandes y pesados.
- **Acceleration**: La aceleración varía entre 8 y 24.8 segundos, con una media de 15.52 segundos, mostrando que el conjunto de datos incluye tanto vehículos deportivos rápidos como modelos más lentos y pesados.
- **Model Year**: Los modelos de vehículos van desde el año 70 hasta el 82, con una media de aproximadamente 75.92, indicando que los datos abarcan un rango de 12 años durante los cuales la industria automotriz experimentó cambios significativos en diseño y tecnología.
- **Origin**: La variable probablemente clasifica los vehículos en categorías basadas en su lugar de fabricación, muestra valores de 1 a 3, con una ligera tendencia hacia el valor más bajo (1.57 de media), lo que podría sugerir una mayor representación de vehículos de un área geográfica específica en el conjunto de datos.

11. Realice un breve análisis exploratorio para identificar la distribución de las variables usadas en la base de datos ¿será que existe relación entre las variables?

11.1. Distribución de las variables

Para analizar la distribución de las variables en el dataset(Ruidoso), se emplearán visualizaciones gráficas, específicamente histogramas. Estas representaciones visuales nos permitirán comprender de manera más clara la dispersión y la forma de cada variable, lo que facilitará la identificación de patrones o características importantes en los datos.

Variables continuas.

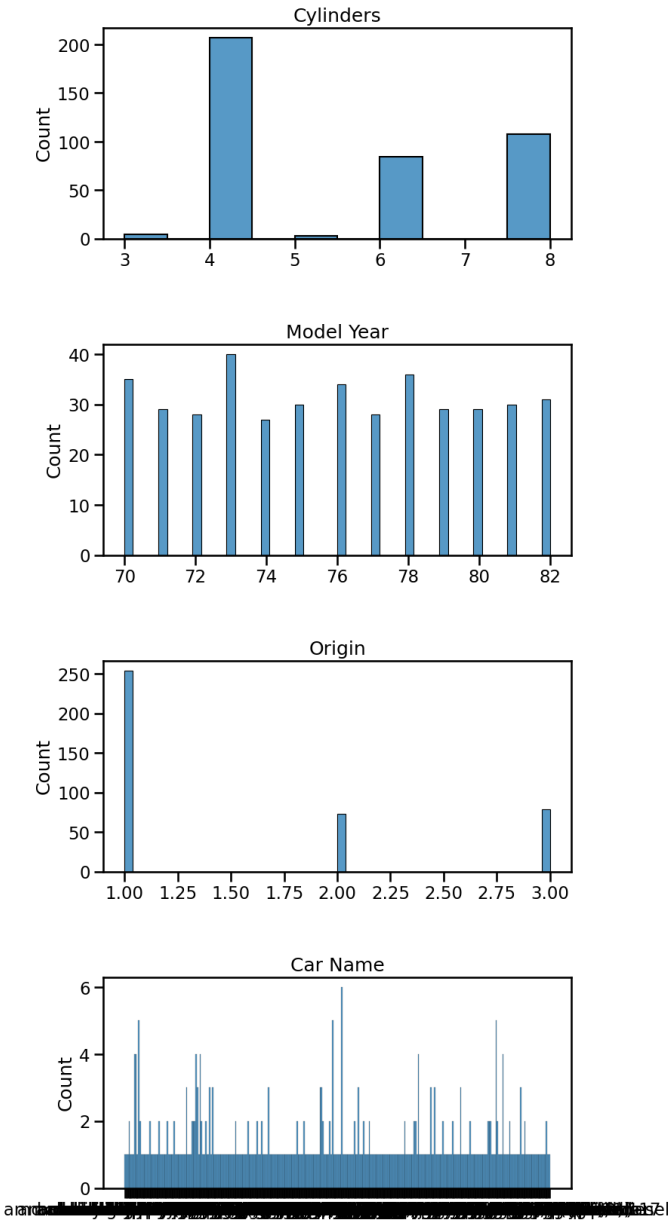


- **MPG** :La mayoría de los vehículos tienen un MPG entre 10 y 30, con el pico más alto alrededor de 15. La distribución parece sesgada hacia la derecha, lo que indica que hay menos vehículos con MPG alto.
- **Desplazamiento** :La distribución parece bimodal, con dos picos distintos, uno alrededor de 100 y otro alrededor de 250. Esto sugiere que hay dos grupos de vehículos con distintas capacidades de motor en la muestra.
- **Horsepower** :El histograma muestra una distribución con un pico principal alrededor de 100 caballos de fuerza, también con un posible sesgo hacia la derecha. Esto indica que, mientras la mayoría de los vehículos tienen un horsepower moderado, hay algunos con valores mucho más altos.
- **Weight** : La distribución de peso de los vehículos muestra un pico alrededor de 3000, con una caída gradual hacia los pesos más altos. Esta distribución también parece

tener un sesgo a la derecha.
- **Acceleration** :La aceleración de los vehículos parece tener una distribución más simétrica con un pico cerca de 15. La forma de la distribución se asemeja a una distribución normal

Obervación:En el caso de la variable "Desplazamiento", observamos una distribución bimodal. Este fenómeno podría ser atribuido a la presencia de vehículos tanto automáticos como manuales en el conjunto de datos, lo cual podría explicar la naturaleza bimodal de la distribución.

Variables int y categoricas:



- **Cylinders** :La variable categórica discreta muestra el número de cilindros en los vehículos. La mayoría tiene 4 cilindros, seguido por un número significativo con 8, y menos con 6. La presencia mínima de vehículos con 3 y 5 cilindros sugiere una preferencia por ciertas configuraciones de motor.
- **Model Year**:Los años del modelo se consideran una variable categórica, ya que cada número representa un año específico en lugar de una cantidad. La distribución es bastante uniforme, con una ligera disminución en los conteos hacia los años más recientes. Esto podría indicar que la base de datos contiene una variedad de modelos de años anteriores, con posiblemente menos datos disponibles para los años más recientes.
- **Origin** :Esta variable parece indicar el país o región de origen de los vehículos. Los valores de 1.00, 2.00 y 3.00 probablemente representan diferentes regiones o países. La mayoría de los vehículos son de la región representada por 1.00, con muchos menos vehículos de las otras dos regiones.
- **Car Name**: Esta es una variable categórica que representa el nombre o el modelo del carro. El histograma muestra una gran cantidad de nombres o modelos únicos

caso variable "Car Name": La variable cuenta con 312 registros únicos, lo que sugiere la necesidad de realizar un proceso de limpieza de datos. Por ejemplo, algunas marcas como "chevorelt" pueden tener versiones automáticas o manuales, pero en esencia representan la misma marca. Automatizar los datos sería beneficioso, ya que trabajar con la información en su estado actual podría ocasionar problemas de segregación. Por lo tanto, se decide dejar la variable sin modificar y no trabajar con ella a menos que el problema lo requiera.

11.2. ¿Cual es la distribución de los datos?

Hay variables continuas pueden distribuirse normalmente, las variables de "Desplazamiento" y "Horsepower" muestran distribuciones con cierto sesgo y no parecen ser normales. La variable "Weight" también parece tener un sesgo hacia la derecha. Sin embargo, la variable "Acceleration" tiene la forma que más se asemeja a una distribución normal, aunque todavía puede haber desviaciones que no se pueden discernir solo con la vista. Para concluir definitivamente si las variables siguen una distribución normal, por lo que se realizara pruebas de normalidad estadística como la prueba de Shapiro-Wilk

11.2.1. Shapiro-Wilk

Table 26. *Resultados de la prueba de Shapiro-Wilk*

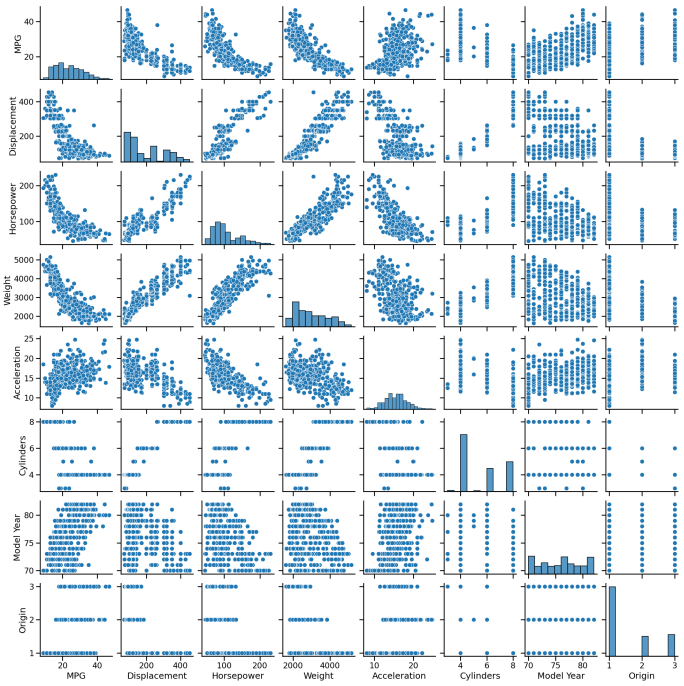
Variable	Estadístico de prueba	Valor p
MPG	0.96	1.05e-07
Displacement	0.88	8.98e-17
Horsepower	0.90	5.02e-15
Weight	0.94	2.60e-11
Acceleration	0.99	0.03

- **MPG**:Se rechaza la hipótesis nula (los datos no provienen de una distribución normal)
- **Desplazamiento**:Se rechaza la hipótesis nula (los datos no provienen de una distribución normal)
- **Horsepower** :Se rechaza la hipótesis nula (los datos no provienen de una distribución normal)
- **Weight**:Se rechaza la hipótesis nula (los datos no provienen de una distribución normal)
- **Acceleration**:Se rechaza la hipótesis nula (los datos no provienen de una distribución normal)

Nota Debemos ser cautelosos con los resultados de la prueba de normalidad mencionada anteriormente, ya que se llevaron

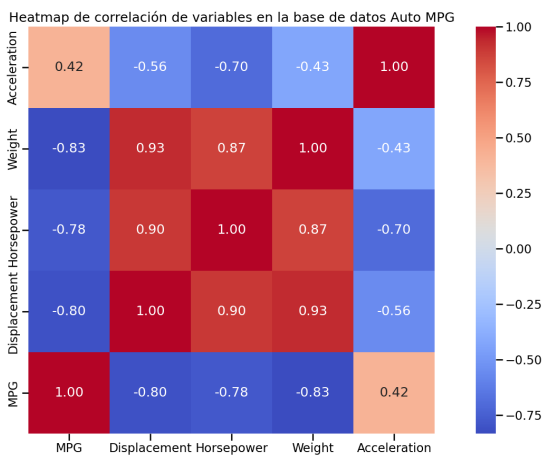
a cabo con los datos después de eliminar los valores NaN. Esto puede enmascarar los resultados, ya que podrían no reflejar fielmente la distribución real de los datos. Una alternativa sería realizar un proceso de imputación para tratar los valores faltantes y luego examinar cómo se comportan los datos en la prueba de normalidad.

11.3. Relación entre variables



Como se discutió en la sección (3.4), es crucial evaluar si la correlación de Pearson es apropiada. Al observar la gráfica, notamos que todas las variables de tipo continuo muestran una relación lineal, ya sea creciente o decreciente. Además, en algunas de ellas se aprecia una ligera curvatura, como en las variables "Horsepower" y "MPG". Por otro lado, se observa que las relaciones entre variables discretas o categóricas, como "Cylinders" y "Model Year", no son apropiadas para la correlación de Pearson. Por lo tanto, en análisis posteriores, se considerará únicamente la correlación entre las variables continuas.

Correlación de pearson para las variables continuas



- : **MPG y Displacement**: Correlación fuertemente negativa -0.80, lo que significa que a medida que el desplazamiento aumenta, el MPG tiende a disminuir.
- **MPG y Horsepower**: También una fuerte correlación negativa (-0.78), indicando que vehículos con más caballos de fuerza suelen tener un menor MPG.
- **MPG y Weight**: Una de las correlaciones más fuertes y negativas (-0.83), sugiriendo que los vehículos más pesados son menos eficientes en términos de consumo de combustible.
- **MPG y Acceleration**: Correlación positiva moderada (0.42), mostrando que vehículos con mejor aceleración tienden a tener un MPG más alto.
- **Displacement, Horsepower, y Weight**: Correlaciones positivas muy fuertes entre ellos (0.90 a 0.93), lo que implica que estos tres atributos aumentan conjuntamente. Vehículos con mayor desplazamiento y caballos de fuerza también tienden a ser más pesados.
- **Acceleration y Weight**: Correlación negativa moderada (-0.43), indicando que vehículos más pesados suelen tener una aceleración más baja.

Estos resultados sugieren que las características de rendimiento del vehículo como el Displacement, Horsepower y Weight están estrechamente interrelacionadas y afectan negativamente la eficiencia del combustible (MPG). La Acceleration, aunque relacionada con la eficiencia del combustible, parece no estar tan fuertemente vinculada con las otras características del vehículo.

Observación: Se consideró realizar la corrección utilizando todas las variables, pero el resultado obtenido fue conforme a las expectativas. Como se mencionó anteriormente, al tratarse de variables categóricas, la correlación de Pearson no identificó ninguna relación significativa. Por lo tanto, no se mencionan estas relaciones. Para explorar la correlación entre estas variables, es necesario recurrir a otros métodos, como la correlación de Spearman o la correlación de Kendall.

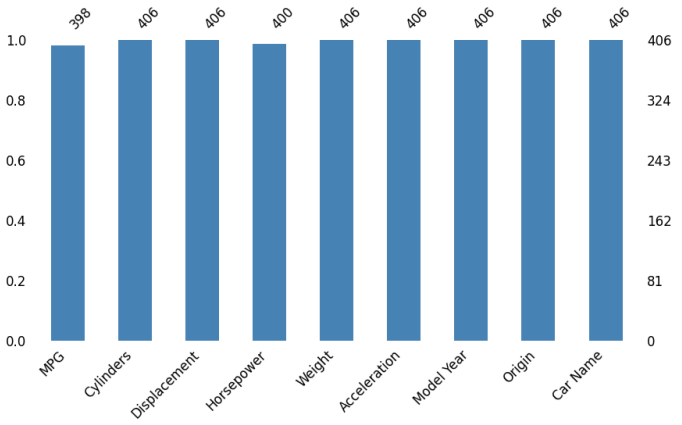
12. Verifique si existen datos faltantes en cada uno de las variables. ¿Cuál es la proporción de datos faltantes en la distribución de las variables?

Como se discutió en la sección (10.1), nuestro enfoque se centrará en la detección de datos faltantes.

12.1. Cantidad de datos faltantes por variable

Table 27. *NaN's*

Variable	Cantidad de NaN's
MPG	8
Cylinders	0
Displacement	0
Horsepower	6
Weight	0
Acceleration	0
Model Year	0
Origin	0
Car Name	0



Podemos observar que únicamente las variables MPG y Horsepower presentan datos faltantes, con 8 y 6 valores faltantes respectivamente.

12.2. Proporción de datos faltantes por variable:

Table 28. *Porcentaje de datos faltantes por variable*

Variable	Porcentaje de datos faltantes
MPG	0.019704 (aproximadamente 1.97%)
Cylinders	0.00(0%)
Displacement	0.00 (0%)
Horsepower	0.014778 (aproximadamente 1.48%)
Weight	0.00 (0%)
Acceleration	0.00(0%)
Model Year	0.00 (0%)
Origin	0.00(0%)
Car Name	0.00 (0%)

En conclusión, la exploración de los datos revela que la mayoría de las variables no tienen valores faltantes. Sin embargo, se observa una pequeña proporción de valores faltantes en las variables de MPG y Horsepower, representando aproximadamente el 1.97% y el 1.48%, respectivamente.

12.2.1. Porcentaje de datos faltantes total

Table 29. *Porcentaje total*

Porcentaje	3.45%

En conclusión, el análisis revela que solo un pequeño porcentaje de datos está ausente en el conjunto de datos analizado, representando aproximadamente el 3.45%. Aunque este porcentaje es relativamente bajo, es importante abordar adecuadamente los valores faltantes para garantizar la integridad y la validez de los análisis posteriores.

13. ¿Cuál cree que es el mecanismo inherente a esos datos faltantes?

La ausencia de datos puede deberse a varios factores, tales como:

- **Facotres propios del procedimento**
  - Formularios mal diseñados
  - Errores de programación
  - etc.
- **Negativa a responder**.Ej:
  - Cuestionarios relacionados con la edad.
  - ¿Cuánto gana?
  - Afilación.
  - Política.
  - Religión.
  - etc.
- **Respuestas inaplicables**.Ej:
  - ¿Cuánto gastos en jugetes para sus hijos el último año?....no tengo hijos!

Ahora, es importante comprender los mecanismos inherentes de los datos, que pueden ser:

- **Outliers tratados como datos faltantes**: Cuando se conocen los limites de las diferentes variables del dataset, los datos que caen fuera del rango definido se deben etiquetar como faltantes.
- **Missing At Random MAR**
  - La probabilidad de que variable **Y** tenga un dato faltante depende de **X**, pero no de **Y**
  - Es decir, el patrón de los datos faltantes se puede predecir a partir de otras variables de la base de datos.
- **Missing completety at random MCAR**
  - La probabilidad de que variables **Y** tenga un dato faltante es independiente de **x**
  - Los datos existentes en **Y** son una muestra al azar de los valores de **y**
- **Non-Ignorable missing data**(Missing noat at random MNAR
  - El valor de la variable que falta relacionado con la razón por la que falta

Ahora cual es mecanismo inherente que explica porque la variable "MPG" y "Horsepower", sabemos que:

- **MPG**: Para esa variable, conocemos la razón de los datos faltantes, ya que en la página donde se descargó el conjunto de datos se explica que "se eliminaron 8 de los casos originales porque tenían valores desconocidos para el atributo 'mpg'". Por lo tanto, podemos inferir que sus valores no concuerdan con el rango de la variable, lo que sugiere que el mecanismo inherente podría ser "**Outliers tratados como datos faltantes**".
- **Horsepower**:Para esta variable, no contamos con ninguna información específica. Por lo tanto, sin datos adicionales, se podría suponer que cada uno de los posibles mecanismos inherentes tendría la misma probabilidad de ocurrencia. Alternativamente, podría ser el mismo caso que para la variable "MPG", donde los datos faltantes podrían haber sido eliminados debido a valores desconocidos o valores atípicos que no concuerdan con el rango de la variable.

14. Aplique las técnicas de tratamiento de datos faltantes vistas en clase.

Los metodos para el tratamiento de datos faltantes son:

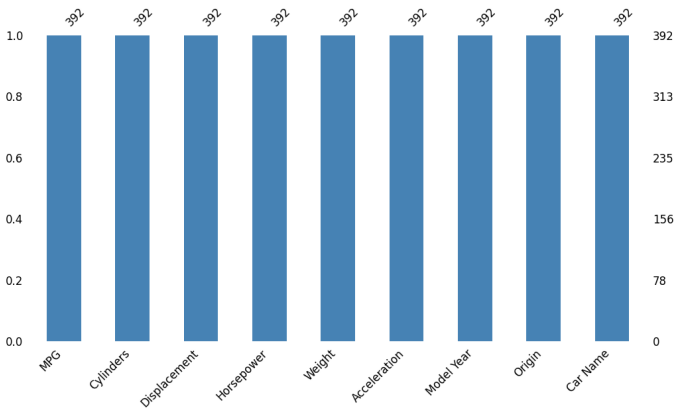
- Método: Eliminación de casos completos.
- Método: "Más frecuente"
- Método: Sustitución por medias
- Método: Cold Deck
- Método: Hot Deck
- Método: Regresión
- Método: MICE
- Método: K-Nearest Neighbor Imputation

14.1. Eliminación de casos completos.

Ya que la cantidad de datos faltantes en ambas variables es inferior a 2%, validemos que tan viable es este método:

Table 30. *NaN's:eliminación de casos completos*

Variable	Cantidad de datos faltantes
MPG	0
Cylinders	0
Displacement	0
Horsepower	0
Weight	0
Acceleration	0
Model Year	0
Origin	0
Car Name	0



	MPG	Cylinders	Displacement	Horsepower
count	392.000000	392.000000	392.000000	392.000000
mean	23.445918	5.471939	194.411990	104.469388
std	7.805007	1.705783	104.644004	38.491160
min	9.000000	3.000000	68.000000	46.000000
25%	17.000000	4.000000	105.000000	75.000000
50%	22.750000	4.000000	151.000000	93.500000
75%	29.000000	8.000000	275.750000	126.000000
max	46.600000	8.000000	455.000000	230.000000

Table 31. *Resumen estadistico. parte 1*

	Weight	Acceleration	Model Year	Origin
count	392.000000	392.000000	392.000000	392.000000
mean	2977.584184	15.541327	75.979592	1.576531
std	849.402560	2.758864	3.683737	0.805518
min	1613.000000	8.000000	70.000000	1.000000
25%	2225.250000	13.775000	73.000000	1.000000
50%	2803.500000	15.500000	76.000000	1.000000
75%	3614.750000	17.025000	79.000000	2.000000
max	5140.000000	24.800000	82.000000	3.000000

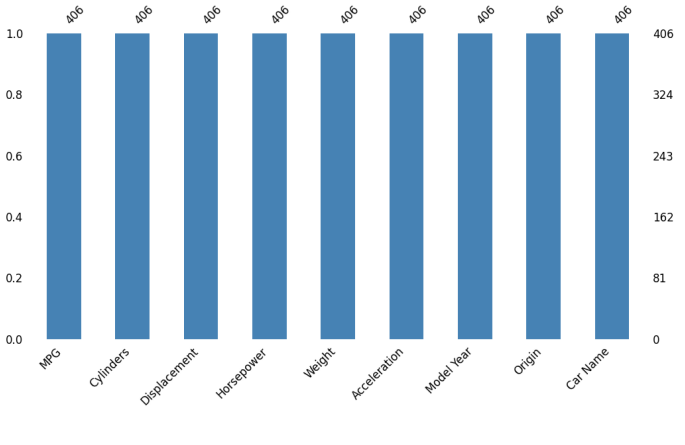
Table 32. *Resumen estadistico. parte 2*

14.2. Más frecuente

Reemplaza los valores faltantes con el valor más frecuente en cada columna.

Table 33. *NaN's: Mas frecuente*

Variable	Cantidad de datos faltantes
MPG	0
Cylinders	0
Displacement	0
Horsepower	0
Weight	0
Acceleration	0
Model Year	0
Origin	0
Car Name	0

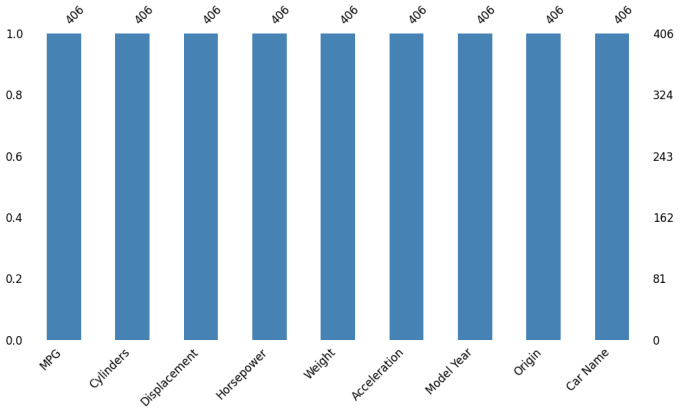


	MPG	Cylinders	Displacement	Horsepower
count	406.000000	406.000000	406.000000	406.000000
mean	23.307389	5.475369	194.779557	105.746305
std	7.875513	1.712160	104.922458	38.861288
min	9.000000	3.000000	68.000000	46.000000
25%	17.000000	4.000000	105.000000	76.000000
50%	22.350000	4.000000	151.000000	95.000000
75%	29.000000	8.000000	302.000000	131.500000
max	46.600000	8.000000	455.000000	230.000000

Table 34. *Resumen estadistico. parte 1*

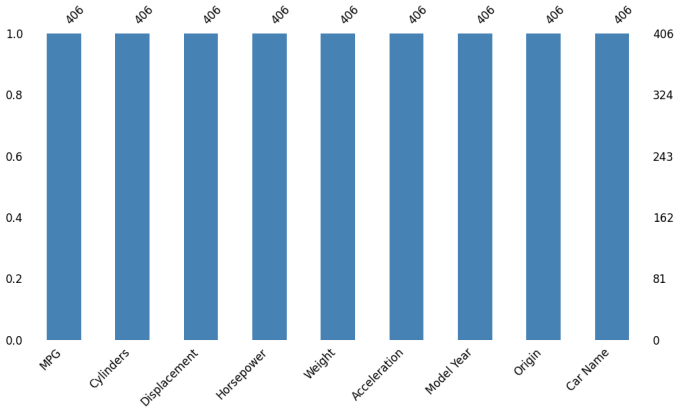
	Weight	Acceleration	Model Year	Origin
count	406.000000	406.000000	406.000000	406.000000
mean	2979.413793	15.519704	75.921182	1.568966
std	847.004328	2.803359	3.748737	0.797479
min	1613.000000	8.000000	70.000000	1.000000
25%	2226.500000	13.700000	73.000000	1.000000
50%	2822.500000	15.500000	76.000000	1.000000
75%	3618.250000	17.175000	79.000000	2.000000
max	5140.000000	24.800000	82.000000	3.000000

Table 35. *Resumen estadistico. parte 2*



14.3. Sustitución por medias

.



	MPG	Cylinders	Displacement	Horsepower
count	406.000000	406.000000	406.000000	406.000000
mean	23.514573	5.475369	194.779557	105.082500
std	7.738404	1.712160	104.922458	38.480531
min	9.000000	3.000000	68.000000	46.000000
25%	17.500000	4.000000	105.000000	76.000000
50%	23.000000	4.000000	151.000000	95.000000
75%	29.000000	8.000000	302.000000	129.000000
max	46.600000	8.000000	455.000000	230.000000

Table 36. *Resumen estadistico. parte 1*

	Weight	Acceleration	Model Year	Origin
count	406.000000	406.000000	406.000000	406.000000
mean	2979.413793	15.519704	75.921182	1.568966
std	847.004328	2.803359	3.748737	0.797479
min	1613.000000	8.000000	70.000000	1.000000
25%	2226.500000	13.700000	73.000000	1.000000
50%	2822.500000	15.500000	76.000000	1.000000
75%	3618.250000	17.175000	79.000000	2.000000
max	5140.000000	24.800000	82.000000	3.000000

Table 37. *Resumen estadistico. parte 2*

14.4. Cold Deck

Utilizaremos la mediana para imputar con el método Cold Deck.

	MPG	Cylinders	Displacement	Horsepower
count	406.000000	406.000000	406.000000	406.000000
mean	23.504433	5.475369	194.779557	104.933498
std	7.738736	1.712160	104.922458	38.499806
min	9.000000	3.000000	68.000000	46.000000
25%	17.500000	4.000000	105.000000	76.000000
50%	23.000000	4.000000	151.000000	95.000000
75%	29.000000	8.000000	302.000000	129.000000
max	46.600000	8.000000	455.000000	230.000000

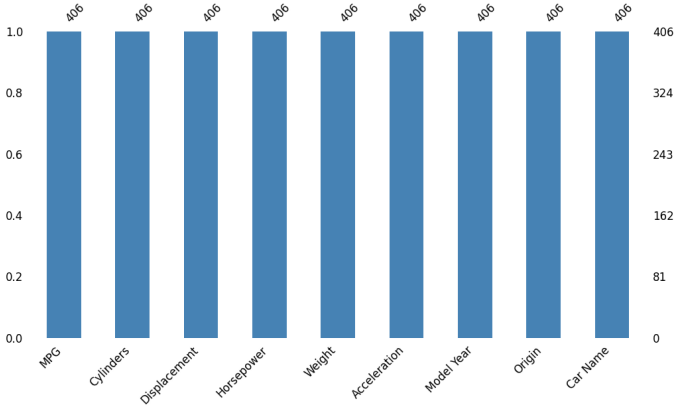
Table 38. *Resumen estadistico. parte 1*

	Weight	Acceleration	Model Year	Origin
count	406.000000	406.000000	406.000000	406.000000
mean	2979.413793	15.519704	75.921182	1.568966
std	847.004328	2.803359	3.748737	0.797479
min	1613.000000	8.000000	70.000000	1.000000
25%	2226.500000	13.700000	73.000000	1.000000
50%	2822.500000	15.500000	76.000000	1.000000
75%	3618.250000	17.175000	79.000000	2.000000
max	5140.000000	24.800000	82.000000	3.000000

Table 39. *Resumen estadistico. parte 2*

14.5. Hot Deck

Para aplicar el método del hot deck para la imputación de valores faltantes, vamos a utilizar un enfoque similar al cold deck, pero en lugar de usar la mediana o el valor más común, vamos a buscar el valor más similar a los registros vecinos.



	MPG	Cylinders	Displacement	Horsepower
count	406.000000	406.000000	406.000000	406.000000
mean	23.258128	5.475369	194.779557	105.349754
std	7.733188	1.712160	104.922458	38.106554
min	9.000000	3.000000	68.000000	46.000000
25%	17.500000	4.000000	105.000000	76.000000
50%	22.000000	4.000000	151.000000	95.000000
75%	29.000000	8.000000	302.000000	130.000000
max	46.600000	8.000000	455.000000	230.000000

Table 40. *Resumen estadistico. parte 1*

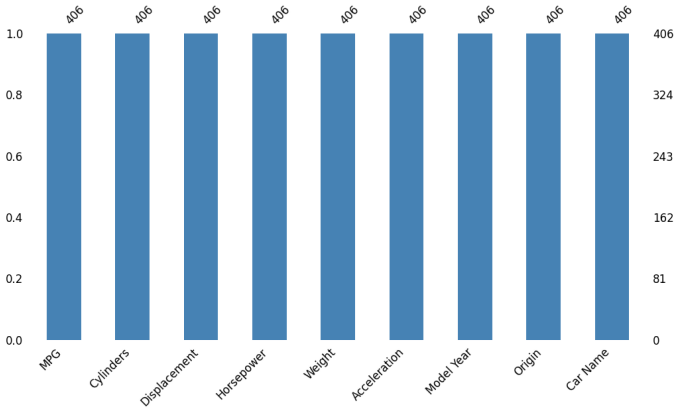
	Weight	Acceleration	Model Year	Origin
count	406.000000	406.000000	406.000000	406.000000
mean	2979.413793	15.519704	75.921182	1.568966
std	847.004328	2.803359	3.748737	0.797479
min	1613.000000	8.000000	70.000000	1.000000
25%	2226.500000	13.700000	73.000000	1.000000
50%	2822.500000	15.500000	76.000000	1.000000
75%	3618.250000	17.175000	79.000000	2.000000
max	5140.000000	24.800000	82.000000	3.000000

Table 41. *Resumen estadistico. parte 2*

14.6. Regresión

creamos un modelo de regresión lineal para predecir los valores faltantes de "MPG" y "Horsepower" en función de las demás variables. Luego, utilizamos este modelo para predecir los valores faltantes y los imputamos en la base de datos.

Se utilizó regresión lineal múltiple, ya que se diseñó un modelo de regresión lineal con múltiples variables predictoras (Cylinders, Displacement, Weight, Acceleration, Model Year) para predecir la variable de interés (MPG o Horsepower) que poseen datos faltantes.



	MPG	Cylinders	Displacement	Horsepower
count	406.000000	406.000000	406.000000	406.000000
mean	23.389240	5.475369	194.779557	104.698211
std	7.811233	1.712160	104.922458	38.711574
min	9.000000	3.000000	68.000000	46.000000
25%	17.000000	4.000000	105.000000	75.128011
50%	22.823893	4.000000	151.000000	94.500000
75%	29.000000	8.000000	302.000000	129.000000
max	46.600000	8.000000	455.000000	230.000000

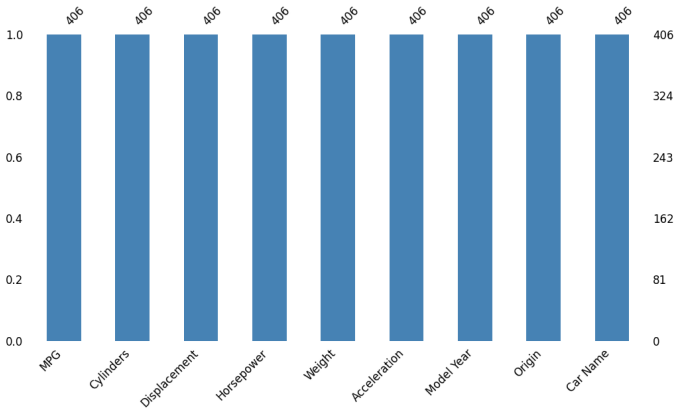
Table 42. *Resumen estadistico. parte 1*

	Weight	Acceleration	Model Year	Origin
count	406.000000	406.000000	406.000000	406.000000
mean	2979.413793	15.519704	75.921182	1.568966
std	847.004328	2.803359	3.748737	0.797479
min	1613.000000	8.000000	70.000000	1.000000
25%	2226.500000	13.700000	73.000000	1.000000
50%	2822.500000	15.500000	76.000000	1.000000
75%	3618.250000	17.175000	79.000000	2.000000
max	5140.000000	24.800000	82.000000	3.000000

Table 43. *Resumen estadistico. parte 2*

14.7. MICE

MICE (Multiple Imputation by Chained Equations) es un método iterativo que utiliza modelos predictivos para estimar los valores faltantes en función de las demás variables.



	MPG	Cylinders	Displacement	Horsepower
count	406.000000	406.000000	406.000000	406.000000
mean	23.417885	5.475369	194.779557	104.815488
std	7.820504	1.712160	104.922458	38.692454
min	9.000000	3.000000	68.000000	36.900219
25%	17.000000	4.000000	105.000000	75.250000
50%	22.547542	4.000000	151.000000	95.000000
75%	29.000000	8.000000	302.000000	129.000000
max	46.600000	8.000000	455.000000	230.000000

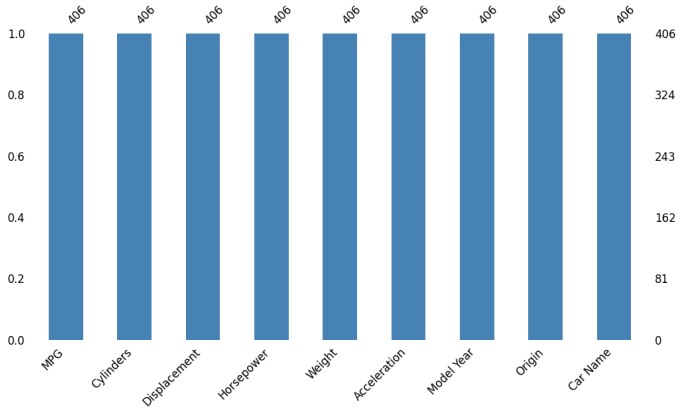
Table 44. *Resumen estadistico. parte 1*

	Weight	Acceleration	Model Year	Origin
count	406.000000	406.000000	406.000000	406.000000
mean	2979.413793	15.519704	75.921182	1.568966
std	847.004328	2.803359	3.748737	0.797479
min	1613.000000	8.000000	70.000000	1.000000
25%	2226.500000	13.700000	73.000000	1.000000
50%	2822.500000	15.500000	76.000000	1.000000
75%	3618.250000	17.175000	79.000000	2.000000
max	5140.000000	24.800000	82.000000	3.000000

Table 45. *Resumen estadistico. parte 2*

14.8. K-Nearest Neighbor Imputation

Imputará las variables "MPG" y "Horsepower" utilizando el método de imputación KNN con 5 vecinos y pesos uniformes.



	MPG	Cylinders	Displacement	Horsepower
count	406.000000	406.000000	406.000000	406.000000
mean	23.429113	5.475369	194.779557	104.779803
std	7.838123	1.712160	104.922458	38.588307
min	9.000000	3.000000	68.000000	46.000000
25%	17.000000	4.000000	105.000000	75.250000
50%	22.750000	4.000000	151.000000	94.900000
75%	29.000000	8.000000	302.000000	129.000000
max	46.600000	8.000000	455.000000	230.000000

Table 46. *Resumen estadistico. parte 1*

	Weight	Acceleration	Model Year	Origin
count	406.000000	406.000000	406.000000	406.000000
mean	2979.413793	15.519704	75.921182	1.568966
std	847.004328	2.803359	3.748737	0.797479
min	1613.000000	8.000000	70.000000	1.000000
25%	2226.500000	13.700000	73.000000	1.000000
50%	2822.500000	15.500000	76.000000	1.000000
75%	3618.250000	17.175000	79.000000	2.000000
max	5140.000000	24.800000	82.000000	3.000000

Table 47. *Resumen estadistico. parte 2*

14.9. Conclusión

Cada uno de estos métodos tiene sus ventajas y aplicaciones específicas, dependiendo de la naturaleza de los datos faltantes y la estructura subyacente del conjunto de datos. Mientras que métodos simples como la eliminación de casos completos o la imputación por el valor más frecuente pueden ser adecuados para conjuntos de datos con pocos valores faltantes y sin preocupaciones sobre la pérdida de información, métodos más complejos como MICE, regresión o KNN son preferibles para conjuntos de datos con patrones subyacentes más complejos y cuando se desea preservar la estructura de datos y las correlaciones entre variables. La elección del método debe guiarse por la comprensión del conjunto de datos, la cantidad y el mecanismo de los datos faltantes, así como los objetivos del análisis.

15. Analice gráfica y analíticamente la variación en la distribución de los datos al aplicar las técnicas de imputación de datos. ¿Qué técnica afecta menos la distribución original?

15.1. Valores MPG de cada metodo

Table 48. *Valores Imputados según Diferentes Métodos*

ID	Más Frecuente	Medias	Cold Deck	Hot Deck
10	13.0	23.514573	23.0	18.0
11	13.0	23.514573	23.0	18.0
12	13.0	23.514573	23.0	18.0
13	13.0	23.514573	23.0	18.0
14	13.0	23.514573	23.0	18.0
17	13.0	23.514573	23.0	18.0
39	13.0	23.514573	23.0	18.0
367	13.0	23.514573	23.0	18.0

Table 49. *Valores Imputados según Diferentes Métodos*

ID	Regresión	MICE	KNN
10	18.360504	21.801565	24.74
11	11.020972	13.866381	14.34
12	11.719783	15.770825	14.48
13	11.022354	12.279344	13.40
14	13.040211	12.279344	13.40
17	15.717557	17.833973	15.50
39	26.622342	32.434712	37.16
367	28.349683	22.595084	20.40

15.2. Valores Horsepower de cada metodo

Table 50. *Valores Imputados para MPG Métodos*

ID	Más Frecuente	Medias	Cold Deck	Hot Deck
38	150.0	105.0825	95.0	130.0
133	150.0	105.0825	95.0	130.0
337	150.0	105.0825	95.0	130.0
343	150.0	105.0825	95.0	130.0
361	150.0	105.0825	95.0	130.0

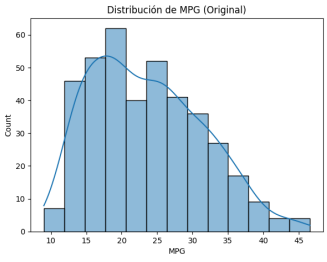
Table 51. *Valores Imputados según Diferentes Métodos*

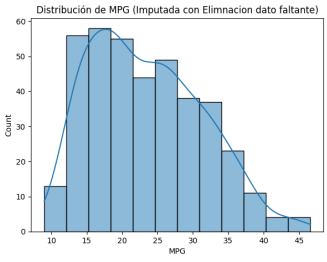
ID	Regresión	MICE	KNN
38	63.910688	98.669219	87.6
133	96.547465	114.208590	93.0
337	57.894505	36.900219	66.6
243	101.990566	104.107999	96.4
361	76.421216	61.763213	69.2

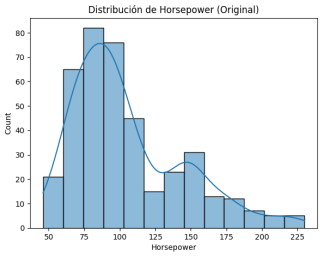
A continuación, procederemos a examinar detenidamente los histogramas acompañados de sus curvas de densidad para evaluar los efectos resultantes de las diversas técnicas empleadas en el tratamiento de datos faltantes.

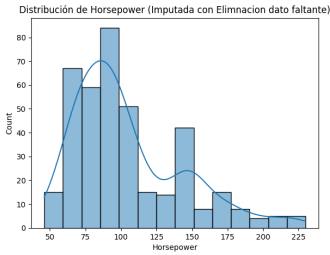
15.3. Histogramas

15.3.1. Método: Eliminación de casos completos

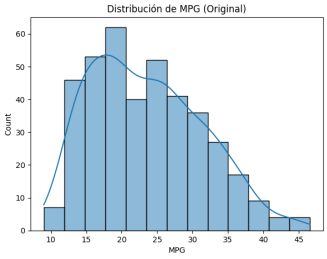


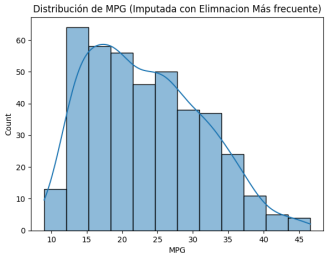


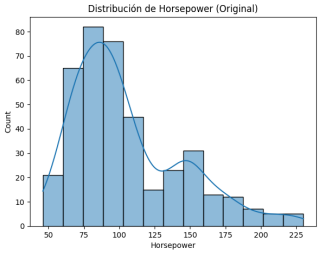


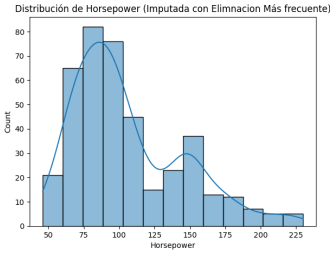


15.3.2. Método: "Más frecuente"

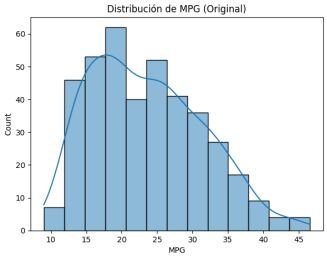


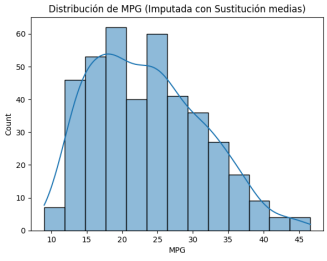


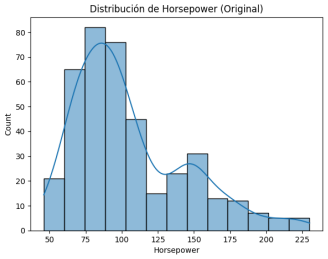


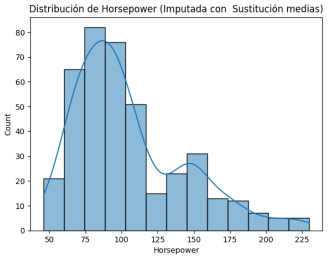


15.3.3. Método: Sustitución por medias

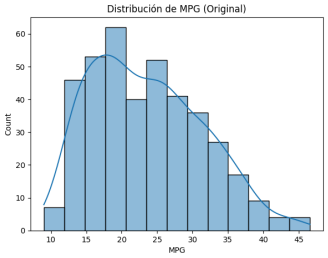


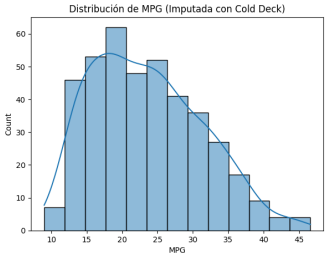


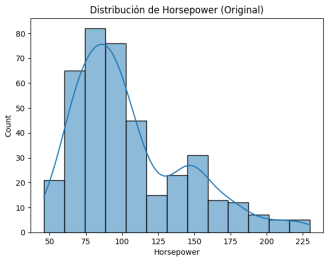


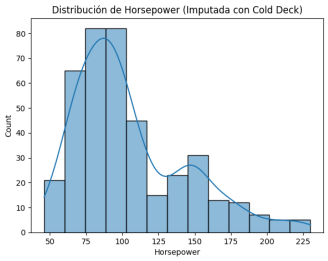


15.3.4. Método: Cold Deck

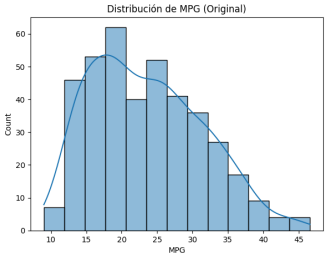


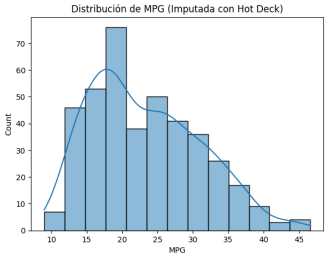


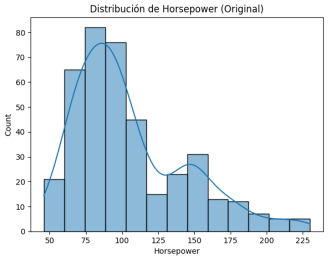


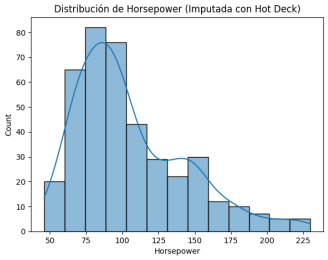


15.3.5. Método: Hot Deck

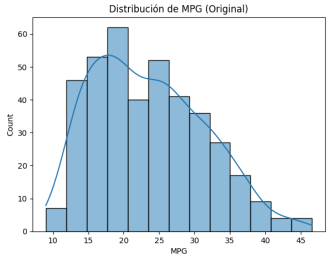


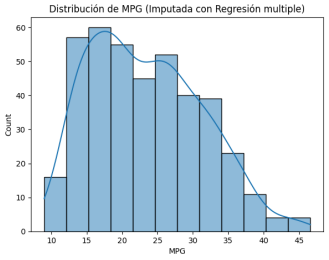


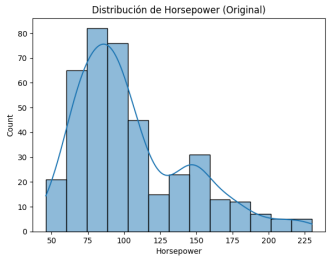


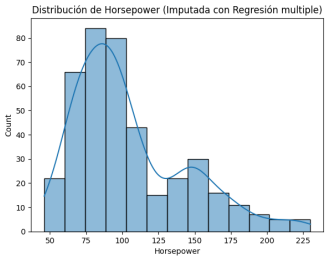


15.3.6. Método: Regresión

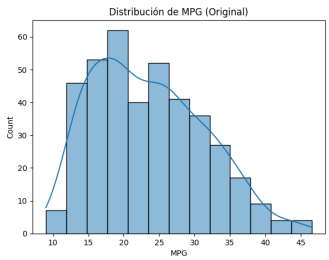


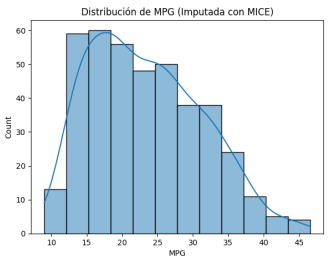


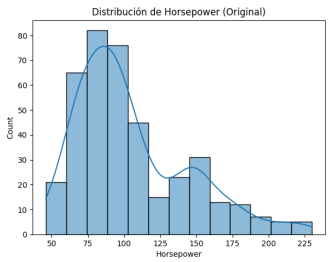


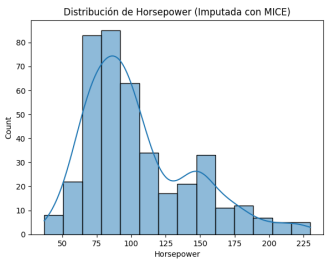


15.3.7. Método: MICE

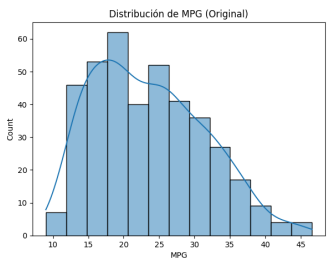


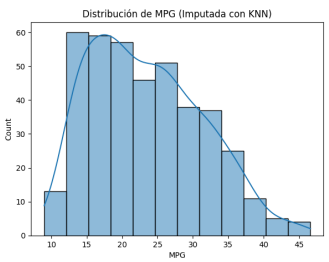


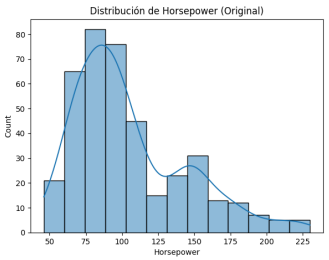


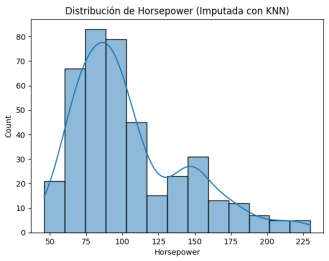


15.3.8. Método: K-Nearest Neighbor Imputation









15.4. ¿Que efecto hay en la normalidad de la variable.

Como se detalla en la **Tabla 26**, sección (11.2.1), se realizó una prueba de Shapiro-Wilk para las variables en cuestión, GPG y Horsepower, cuyos resultados indicaron que no seguían una distribución normal. Surge la pregunta de si los métodos de imputación podrían influir en su distribución. Además, en la mencionada sección se incluyó una nota aclaratoria sobre cómo la ausencia de datos completos podría afectar los resultados de las pruebas, ya sea que se realicen con o sin los datos faltantes. Por lo tanto, después de aplicar los procedimientos de imputación, examinaremos si hay cambios en la distribución. A priori, se podría pensar que no habrá cambios significativos, dado que la proporción de valores faltantes en estas variables es muy baja, como se evidenció en la Tabla 28, sección 12.2.

15.4.1. Shapiro-wilk para el metodo:Sustituciónn de medias

Table 52. *Resultados de la prueba de Shapiro-Wilk*

Variable	Estadístico de prueba	Valor p
MPG	0.97	1.06e-07
Horsepower	0.91	9.825e-15

- **MPG**:Se rechaza la hipótesis nula (los datos no provienen de una distribución normal)
- **Horsepower** :Se rechaza la hipótesis nula (los datos no provienen de una distribución normal)

Como se puede observar, los valores p experimentan ciertas variaciones en comparación con los obtenidos previamente; sin embargo, la conclusión general permanece inalterada. Esto sugiere que la aplicación del método de imputación no altera significativamente las distribuciones de las variables analizadas. Este resultado se alinea con la observación previa de que la cantidad de datos faltantes no es suficientemente grande como para impactar de manera significativa los resultados de las pruebas estadísticas.

Nota: Es importante destacar que esta conclusión no es exclusiva del método de imputación mencionado; se ha aplicado el mismo análisis a otras variables y métodos, obteniendo resultados similares. La decisión de enfocarse en un solo ejemplo fue tomada para simplificar la exposición del impacto de la imputación en la distribución de los datos. Sin embargo, se aseguró que todos los análisis necesarios fueron llevados a cabo para validar esta conclusión dada.

15.5. Conclusión

El análisis de los métodos de imputación sobre la presente base de datos revela que, dada la modesta proporción de datos faltantes (aproximadamente un 4%), las técnicas empleadas no inducen alteraciones significativas en las características fundamentales de las variables analizadas. Se constata que las medias de las variables se mantienen estables, sin experimentar cambios notables. Además, se observa que la mayoría de los métodos tienden a "suavizar" la distribución de los datos, facilitando una cierta consistencia sin distorsionar excesivamente su estructura original. En particular, para la variable "MPG" (Millas por galón), se destaca que estrategias como la sustitución por la media y el método cold deck conservan eficazmente la distribución original de los datos, evidenciando solo ajustes mínimos en medidas de tendencia central y dispersión, e incluso registrando mejoras marginales en la desviación estándar. Contrariamente, técnicas más elaboradas muestran limitaciones para replicar con fidelidad la distribución original, distanciándose en ciertos casos de las características iniciales de los datos.

Por otro lado, con respecto a la variable "Horsepower" (Caballos de fuerza), la influencia de los métodos de imputación es relativamente menor, logrando en su mayoría preservar las cualidades esenciales de la distribución de los datos, sin afectar de manera significativa estadísticas descriptivas clave como la media, la mediana o la varianza. Finalemente, este estudio demuestra que, para el conjunto de datos en cuestión, es viable la aplicación de métodos de imputación relativamente sencillos y de baja complejidad computacional. Estas técnicas no solo facilitan la gestión de los datos faltantes sino que también aseguran la preservación de la integridad y la estructura original de la base de datos, permitiendo una interpretación y análisis confiables.

16. Apendice

En el siguiente código QR, encontrarás acceso al código base utilizado para realizar todos los análisis presentados en este trabajo:

Si encuentras dificultades, puedes acceder mediante el siguiente enlace: https://drive.google.com/file/d/1jFKP9dtl-M$_2$*umwdTCgeLNlPGTcwyYp*1/*view*?*usp* = *sharing* Además, el codigo se enviará de forma separada junto con este trabajo.

17. Referencias

- UCI Machine Learning Repository. (n.d.). Auto MPG Data Set. University of California, Irvine. Recuperado el [21/03/24] de https://archive.ics.uci.edu/ml/datasets/auto+mpg
- Facultad de Ciencias Exactas, Ingeniería y Agrimensura, Universidad Nacional de Rosario NIVELES SONOROS. Recuperado de https://www.fceia.unr.edu.ar/acustica/comite/niveles.htm
- Zamorano González, B., Peña Cárdenas, F., Parra Sierra, V., Velázquez Narváez, Y., & Vargas Martínez, J. I. (2015). Contaminación por ruido en el centro histórico de Matamoros. Acta universitaria, 25(5), 20-27.
- Guarin Escudero, J. V. (2024). Semana 5 2024 01.pdf [Notas de clase].Universidad Nacional de colombia

Contact:
Jsandovalma@unal.edu.co
jgarciaza@unal.edu.co
kagomezc@unal.edu.co