Archivo: Análisis Exploratorio de datos (EDA)

Autor: Castillo Gomez Carlos Alexander

Materia: Introducción a la Ciencia de datos

Profesor: Jaime Alejandro Romero Sierra

Fecha de entrega: 25/11/2024

Análisis Exploratorio de datos (EDA)

1. Descripción General de los datos.

Visión General. El dataset incluye un total de 18 columnas con 4618 filas, el dataset tiene 6 columnas (Marca, Submarca, Versión, Trans., Comb. Y Categoría) de un tipo object, es decir, esas columnas en su mayoría son texto. 6 columnas (Modelo, Cilindros, Potencia (HP), CO2(g/km), NOx (g/1000km) y Calificación Gas Ef. Inv.) de tipo int64, donde podemos encontrar datos numéricos pero enteros. Finalmente, las otras 6 columnas (Tamaño (L), R. Ciudad (km/l), R. Comb. (km/l), R. Ajust. (km/l), R. Carr. (km/l) y Calificación Contam. Aire) son de tipo float64, en ellas podemos encontrar datos numéricos pero que pueden llegar a incluir decimales.

Tipos de Variables. Las variables como ya lo había mencionado son 3, tipo object donde encontramos información sobre el automóvil en forma de texto, la marca es un ejemplo. También tenemos 2 variables numéricas que estas son int y float, las primeras nos muestran datos enteros positivos, como la potencia mientras que las de float nos muestran los datos más estadísticos como pueden ser el rendimiento del automóvil de acuerdo con el lugar donde se encuentre.

Resumen estadístico. Gracias a la función describe (), pudimos obtener de una forma más eficaz el promedio de las columnas de tipo numérico, donde podemos apreciar que el promedio del modelo de los autos es 2014, pero en especial se pueden apreciar los datos del promedio de la Calificación para Gas de efecto invernadero y Calificación para contaminación del aire donde tenemos un 4.88 y 7.89 respectivamente, lo cual si es preocupante pues no son promedios que se escuchen del todo bien, son muy bajos.

Mientras que en la mediana de nuestros datos se aprecia que los autos en México la mayoría son de 4 cilindros, además que la mediana de la potencia de la mayoría de los autos que circulan en el país es de 220.

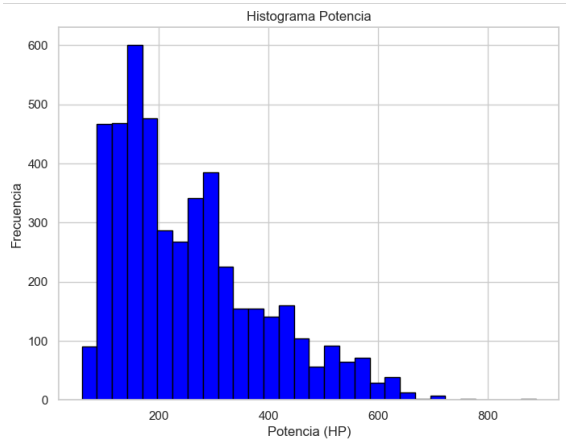
	Modelo	Cilindros	Potencia (HP)	Tamaño (L)	R. Ciudad (km/l)	R. Carr. (km/l)	R. Comb. (km/l)	R. Ajust. (km/l)	CO2(g/km)	NOx (g/1000km)	Calificación Gas Ef. Inv.	Calificación Contam. Aire
count	4692.000000	4692.000000	4692.000000	4692.000000	4692.000000	4692.000000	4692.000000	4692.000000	4692.000000	4692.000000	4692.000000	4692.000000
mean	2014.185422	5.319480	254.224851	2.864214	10.613645	16.625335	13.202383	9.897226	256.218883	30.941816	4.885337	7.895455
std	2.153494	1.790218	132.302071	1.340479	3.288468	4.185040	3.608934	2.701238	75.401836	57.673629	2.479948	1.219092
min	2011.000000	3.000000	60.000000	0.900000	3.100000	6.700000	4.960000	3.720000	107.000000	0.000000	0.000000	1.000000
25%	2012.000000	4.000000	150.000000	1.800000	8.200000	13.500000	10.490000	7.880000	200.000000	10.000000	3.000000	7.895356
50%	2014.000000	4.000000	220.000000	2.500000	10.500000	16.400000	12.930000	9.695000	244.000000	17.000000	5.000000	7.895356
75%	2016.000000	6.000000	327.000000	3.600000	12.812500	19.600000	15.622500	11.702500	298.000000	28.000000	7.000000	9.000000
max	2018.000000	12.000000	888.000000	8.400000	27.460000	31.300000	28.930000	21.700000	627.000000	724.000000	10.000000	9.000000

```
La mediana de la columna 'Cilindros' es: 4.0
```

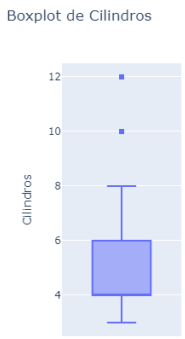
```
La mediana de la columna 'Potencia (HP)' es: 220.0
```

2. Visualización y Distribución de Variables Individuales

Variables numéricas. Después de analizar las columnas se pudo apreciar que existían columnas con outliers, lo cual se detectó con la realización de boxplots, que como se comentaban puntos atrás, las columnas donde se detectó esto fueron corregido con ayuda de códigos para que así los boxplots quedaran “Limpios” y no tuviéramos la posibilidad de sesgos.

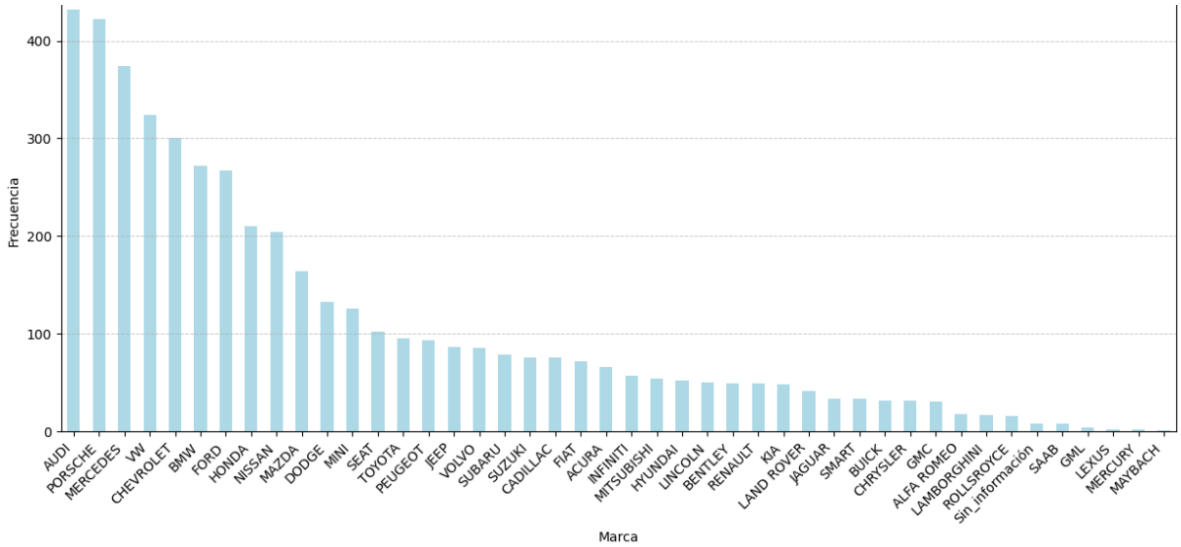


Ahora tenemos un histograma que para ser más específicos corresponde a la columna de Potencia (HP), al verla podemos notar que la mayoría de nuestros datos están en un rango de (30-650) aproximadamente, aunque podemos notar que existen unos ligeros datos después del 650, por lo que son datos que son necesario tomar en cuenta para saber qué hacer con ellos,



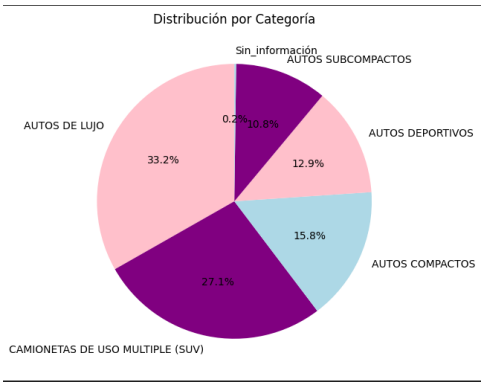
Ahora tenemos este Boxplot que pertenece a la columna cilindros, en ella rápidamente apreciamos 2 puntos en 12 y 10, esto principalmente a que en México como anteriormente se comentó al sacar la mediana y promedio, los autos que circulan en su mayoría son de 4 cilindros, haciendo que los de 10 y 12 sean un lujo, pues principalmente se sabe que entre más cilindros mas consumo de gasolina se genera además que los autos que poseen esta gran cantidad de cilindros en su mayoría son de un tipo superdeportivo, que muy pocos en el país son capaces de adquirir.

Variables Categóricas. Con ayuda de la creación de gráficos de barras, se encontró bastante información.



Primero, con ayuda de esta grafica de barras que pertenece a la columna Marca, podemos observar que la marca Audi es la que más prevalece en nuestro dataset, mientras que marcas más costosas como Lamborghini, RollsRoyce, Alfa Romeo,

etc. Son las que aparecen en menor cantidad en nuestro dataset, esto se puede deber a muchos factores, uno es el precio de estos autos además de que son marcas no tan conocidas, ya que si preguntamos a la población sobre marcas de autos probablemente nos respondan más personas Audi, a una marca como podría ser Maybatch.

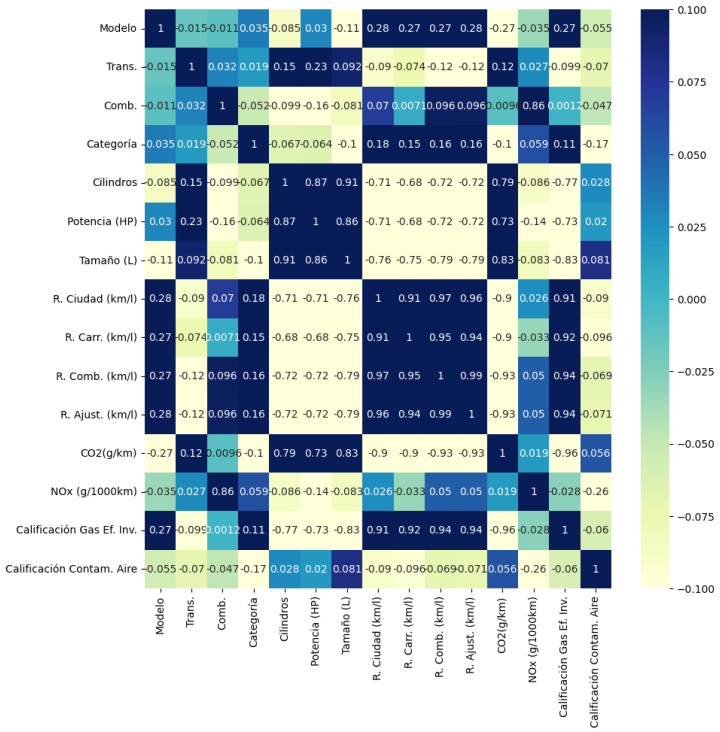


Ahora con esta grafica de pastel notamos que en México en la categoría de auto que más hay en el país son la de autos de lujo, que bien, a pesar de que la palabra “Lujo” sea signo de caro, la mayoría de los autos son conocidos como lujo al incluir cosas que en otros no, aunque estas no sean tan costosas, unos ejemplos son los quemacocos, vidrios eléctricos, pantalla, volante con botones, etc. La mayoría de los autos empiezan a tener estas características por lo que no es novedad que este esa categoría en primer lugar.

3. Correlación entre variables

Matriz de correlación.

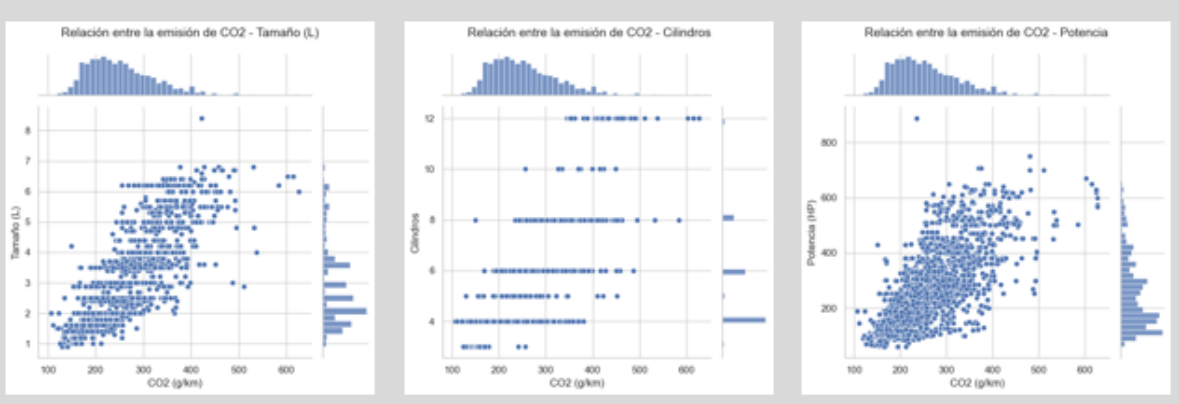
	Modelo	Trans.	Comb.	Categoría	Cilindros	Potencia (HP)	Tamaño (L)	Ciudad (km/l)	R. Carr. (km/l)	R. Comb. (km/l)	R. Ajust. (km/l)	CO2(g/km)	NOx (g/1000km)	Calificación Gas Ef. Inv.
Modelo	1.000000	-0.014930	-0.011140	0.035282	-0.085207	0.029787	-0.106898	0.275665	0.274931	0.273851	0.276575	-0.270076	-0.034924	0.268027
Trans.	-0.014930	1.000000	0.032247	0.019213	0.146395	0.229084	0.091757	-0.090016	-0.074319	-0.119875	-0.118949	0.115311	0.027417	-0.098686
Comb.	-0.011140	0.032247	1.000000	-0.052134	-0.098745	-0.156725	-0.081494	0.069985	0.007095	0.096407	0.096084	-0.009560	0.856357	0.001206
Categoría	0.035282	0.019213	-0.052134	1.000000	-0.066770	-0.064239	-0.104157	0.180269	0.145581	0.160086	0.159302	-0.104707	0.058976	0.111871
Cilindros	-0.085207	0.146395	-0.098745	-0.066770	1.000000	0.865013	0.909346	-0.710240	-0.683880	-0.721914	-0.723627	0.794758	-0.086346	-0.769174
Potencia (HP)	0.029787	0.229084	-0.156725	-0.064239	0.865013	1.000000	0.859222	-0.708913	-0.675592	-0.722499	-0.723011	0.731051	-0.136992	-0.728144
Tamaño (L)	-0.106898	0.091757	-0.081494	-0.104157	0.909346	0.859222	1.000000	-0.758141	-0.745447	-0.785040	-0.785373	0.829864	-0.082686	-0.828127
R. Ciudad (km/l)	0.275665	-0.090016	0.069985	0.180269	-0.710240	-0.708913	-0.758141	1.000000	0.905232	0.965371	0.963060	-0.896029	0.026169	0.912913
R. Carr. (km/l)	0.274931	-0.074319	0.007095	0.145581	-0.683880	-0.675592	-0.745447	0.905232	1.000000	0.945086	0.944082	-0.901441	-0.033247	0.917643
R. Comb. (km/l)	0.273851	-0.119875	0.096407	0.160086	-0.721914	-0.722499	-0.785040	0.965371	0.945086	1.000000	0.992875	-0.925796	0.050008	0.942317
R. Ajust. (km/l)	0.276575	-0.118949	0.096084	0.159302	-0.723627	-0.723011	-0.785373	0.963060	0.944082	0.992875	1.000000	-0.925030	0.049564	0.941221
CO2(g/km)	-0.270076	0.115311	-0.009560	-0.104707	0.794758	0.731051	0.829864	-0.896029	-0.901441	-0.925796	-0.925030	1.000000	0.018887	-0.956496
NOx (g/1000km)	-0.034924	0.027417	0.856357	0.058976	-0.086346	-0.136992	-0.082686	0.026169	-0.033247	0.050008	0.049564	0.018887	1.000000	-0.028443
Calificación Gas Ef. Inv.	0.268027	-0.098686	0.001206	0.111871	-0.769174	-0.728144	-0.828127	0.912913	0.917643	0.942317	0.941221	-0.956496	-0.028443	1.000000



Esta imagen corresponde a un heatmap, en el podemos apreciar variables numéricas, ya que anteriormente se realizaron cambios en las columnas para que sean consideradas números. La importancia que debemos tener en cuenta es la columna CO2 (g/km) pues lo que estamos buscando es que es lo que principalmente influye que los autos generen las emisiones de este gas, para saber eso, revisamos la fila CO2 (g/km) y revisamos sus choques que tiene con cada columna, mientras el tono de azul sea más fuerte es porque es más la relación que tiene esa columna con el gas, notamos que las 3 principales columnas que tomaremos a investigar son los cilindros, potencia y tamaño.

Ya que tenemos las columnas, se llegó a la conclusión que si se tiene relación estas columnas con la emisión de CO2, un ejemplo es la potencia pues entre mas potencia genere más combustible se necesita, que el combustible como sabemos es quien genera este gas.

Parejas de Variables.



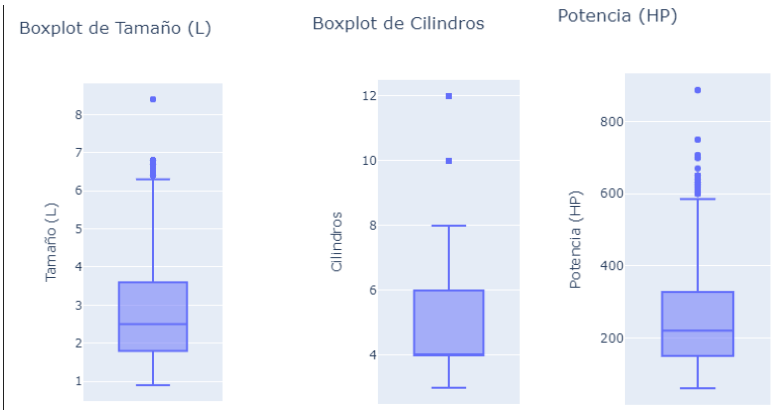
Ahora tenemos gráficos de dispersión, estos nos ayudan a entender de mejor forma como es que se relacionan las columnas ya mencionadas con el CO2, podemos observar que todas entre mayor sea tamaño, potencia o los cilindros mas generación de CO2 se da.

Con esto podemos estar más seguro de hipótesis que al principio se realizaron, y es que si, entre mas grande sea el auto, probablemente sea mayor la potencia que necesite pues no sabemos el uso del automóvil, y por ende mayor serán los cilindros que se ocupen.

Tomemos de ejemplo una camioneta para uso de carga, los cilindros son mayores pues necesita tener una potencia para cargar pesado y como sabemos las camionetas son de un tamaño mayor.

4. Análisis de valores Atípicos (Outliers)

Identificación de Outliers. Como comentábamos anteriormente, se utilizaron boxplots para que encontrara los datos atípicos que existieran en la dataset,

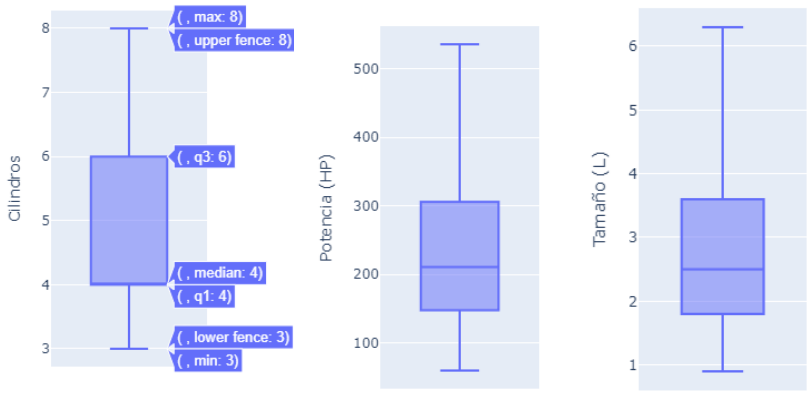


Tomemos como ejemplo este boxplot que pertenece a la columna tamaño, columna que se sabe que se relaciona con lo que buscamos que es el CO2.

Podemos notar que existen valores atípicos en un rango de (6.5 – 8.5) aproximadamente.

Eso notamos también en las otras columnas que se relacionan a lo que estamos buscando, por lo que será necesario realizar algo con estos datos que se especificará más a fondo en el siguiente punto.

Tratamiento de Outliers. En el paso anterior se encontró que si hay datos atípicos en nuestras variables que ocupamos. La mayoría de los datos eran valores que superan los 1.5x el rango intercuartil en los boxplots, por lo que con ayuda de los códigos se eliminaran, la decisión del porque se eliminaron fue que son muy pocos datos que están fuera, pues la mayoría se encuentran en rangos mas grandes, haciendo que yo crea que es mejor investigar sobre esos rangos pues eso nos quiere decir que la mayoría de la población tiene automóviles con características similares, haciendo que investigar sobre los datos atípicos sea algo innecesario teniendo en cuenta que muy pocas personas tienen autos con dichas características.

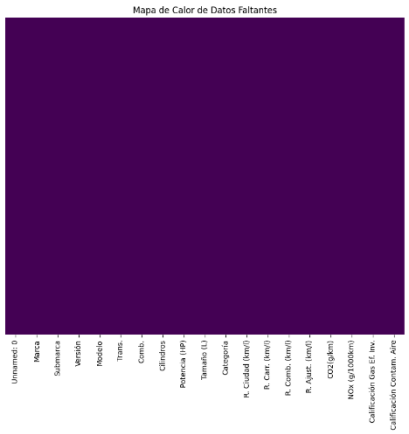


5. Análisis de valores faltantes

Identificación de datos faltantes. Se realizo un mapa de calor de datos ausentes, el cual no informo nada ya que solo apareció un bloque de color y es que si bien, en el pasado se realizo una limpieza de datos donde se eliminaron datos duplicados, se cambiaron datos para así evitar valores inválidos o nulos, no fue necesario realizar alguna estrategia de imputación o eliminación pues al final de cuentas este tipo de datos fueron tratados anteriormente en la limpieza del dataset.

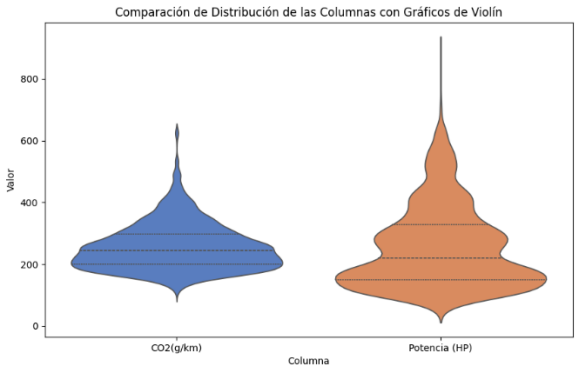
En el pasado algunos datos nulos o faltantes fueron eliminados y otros fueron cambiados a la palabra “Sin_información”.

```
# Cargar el dataset (puedes cargarlo desde un archivo CSV)
df9 = pd.read_csv('Base_limpia_proyecto_CACG')  # Cambia 'ruta_del_archivo.csv' por el archivo que desees

# Crear un mapa de calor de los valores faltantes
plt.figure(figsize=(10, 8))
sns.heatmap(df9.isnull(), cbar=False, cmap='viridis', yticklabels=False, xticklabels=True)

# Títulos y mostrar el mapa
plt.title('Mapa de Calor de Datos Faltantes')
plt.show()
```



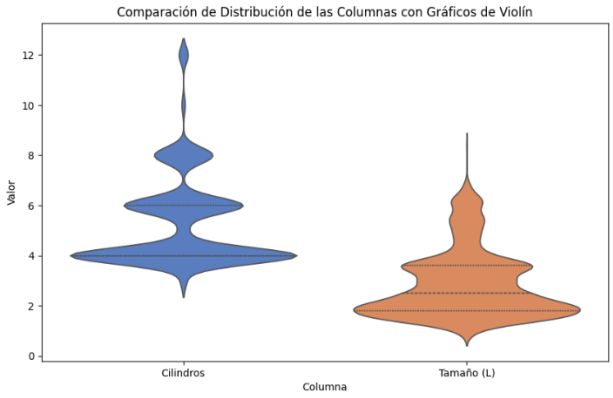
6. Relación entre variables categóricas y numéricas

Análisis comparativo.



Tenemos ahora esta comparación entre 2 graficas de violín, en este caso se puede observas que el top de generación de CO2 se encuentra por los 200-300 aproximadamente mientras que potencia está un poco debajo del 200, por lo que si tendrán relación estas columnas como se dijo anteriormente pero no por eso comparten los mismos rangos, es decir la potencia no determina cuanto CO2 genera un auto.

Y así pasa con cada una de las columnas que tenemos que investigar.



7. Observaciones y hallazgos importantes.

Finalmente, este es un resumen de lo que se encontró en el paso 3 (Metodología).

- La generación de CO2 por parte de los autos se debe principalmente a los cilindros, la potencia, y el tamaño del vehículo.

- Entre mayor sea alguna de las 3 variables que se relacionan con este gas, mayor será la generación de este.
- Me parece interesante que los gráficos tengan un patrón similar haciendo una línea diagonal de forma ascendente.
- Los autos de lujo son más presentes que autos compactos.
- La mayoría de la población tiene autos con características similares.
- No existió ninguna anomalía al parecer.

Estos datos pueden influir en nuestro modelo de machine learning que será “Regresión lineal múltiple” al saber que estos van en ascendencia por lo que espero y se siga cumpliendo este tipo de grafica en el modelo, pues al parecer entre más sea de algo, más subirá el CO2.