Nombre del proyecto: Autos en México y el CO2

Autor: Castillo Gomez Carlos Alexander

Materia: Introducción a la Ciencia de datos

Profesor: Jaime Alejandro Romero Sierra

Fecha de entrega: 25/11/2024

Introducción.

Descripción.

El objetivo de este proyecto es conocer mas a fondo por qué los automóviles son parte de los principales generadores de CO2, que como sabemos este gas ha sido signo de preocupación durante estos últimos años pues contamina nuestro aire.

Justificación y contexto.

El CO2 es un gas incoloro, inodoro y compuesto por oxígeno y carbono. Este gas es de las principales causas del calentamiento global, un problema que ha sido causado por las actividades que el ser humano realiza cotidianamente y agravado por la larga pervivencia del CO2 en la atmósfera.

Una de las principales razones de el porque se han multiplicado las emisiones de este gas es el uso excesivo del automóvil, pues hoy en día el automóvil es una necesidad a comparación del pasado que solo eran pocos los que tenían el lujo de adquirir una unidad. Hoy se ocupa para la mayoría de nuestras actividades cotidianas, y debido a factores del automóvil es el porque se genera CO2, haciendo que sea algo alarmante pues cada día nuestro aire esta mas contaminado y en el futuro esto tendrá consecuencias.

Por lo que es necesario hacer conciencia de nuestras acciones ya que al final es el lugar donde todos vivimos, tenemos que cuidarlo antes de que sea tarde.

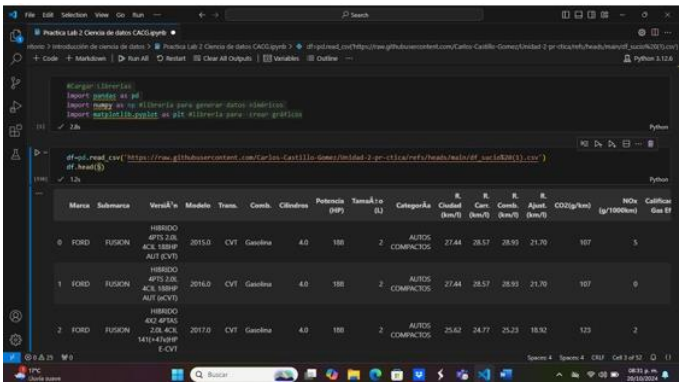
Fuentes de datos.

La base de datos que fue utilizada para la investigación y apoyo de este proyecto fue obtenida de la plataforma Kaggle, propiedad de Alex Acosta quien compartió esta base hace 7 años, en ella encontramos datos relacionados con la marca, submarca, modelo y diferentes características del motor de los diferentes automóviles que solemos ver transitando en las calles de México. Esta base de datos contiene 18 columnas y 4618 filas, al abrirla encontramos marcas como Ford,

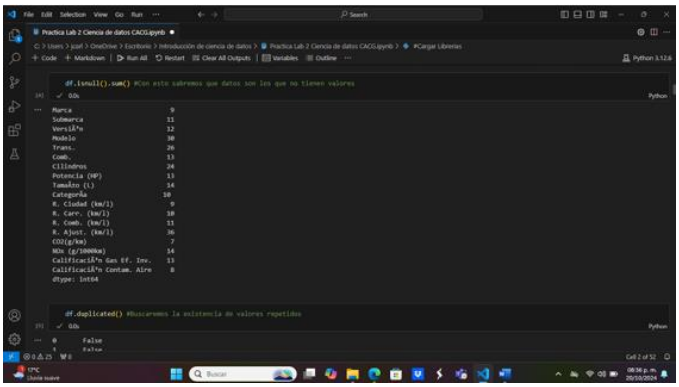
Audi, Seat, etc. Además de información que nos ayuda más a entender el como los autos se clasifican, además de la generación de CO2 de cada uno.

Metodología

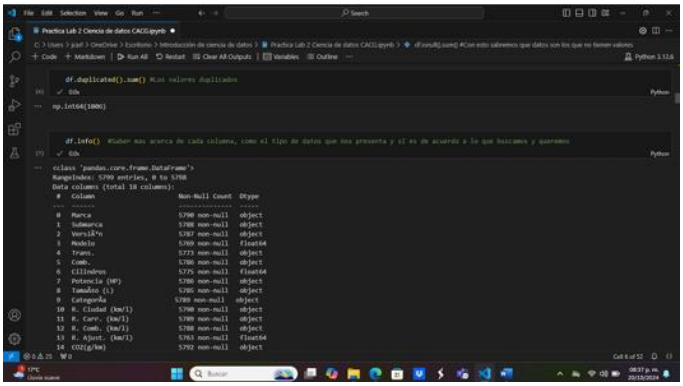
1. Comencé con cargar las librerías, las que ocupe fue pandas y matplotlib, ya después de eso cargue en Visual code el link de la base de datos que subí a Github.



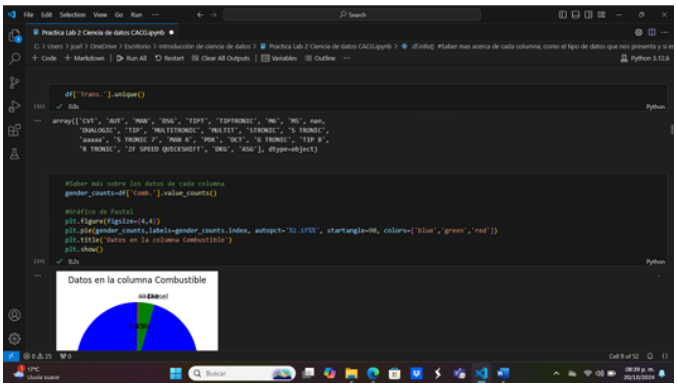
2. Ya comenzando el análisis de las principales funciones que ocupe fue el “isnull (). sum ()” y “duplicated” esto para saber cuántos datos no tenían valores y para buscar la existencia de valores repetidos.



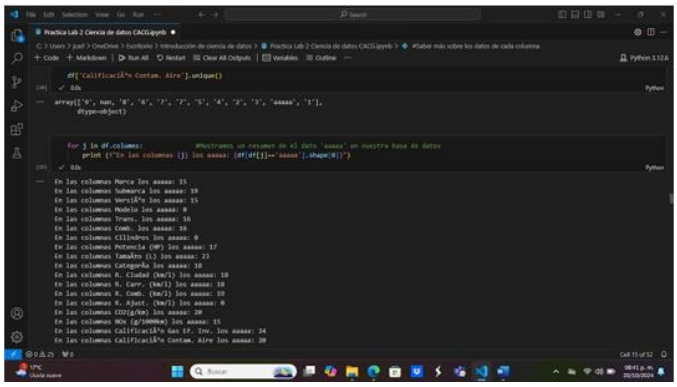
3. También ocupe la función “duplicated (). sum () para saber los valores duplicados y finalmente el “info ()” para saber más acerca de cada columna y principalmente que tipo de datos tenía cada una



4. Con funciones como “unique” e incluso la creación de gráficas pude saber mas acerca de que datos eran los que existían en cada columna, de ahí pude saber que datos que no concordaban era “nan” y “aaaaa”.



5. Por lo que hice eso con cada columna de acuerdo a cuantos datos presentaba era la función que ocupé y al final saber cuántos “aaaaa” existían en cada columna.

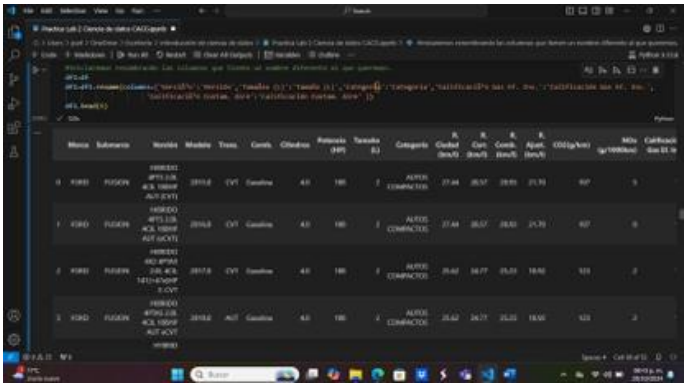


6. Ya después de analizar los “errores” realizar la conclusión.

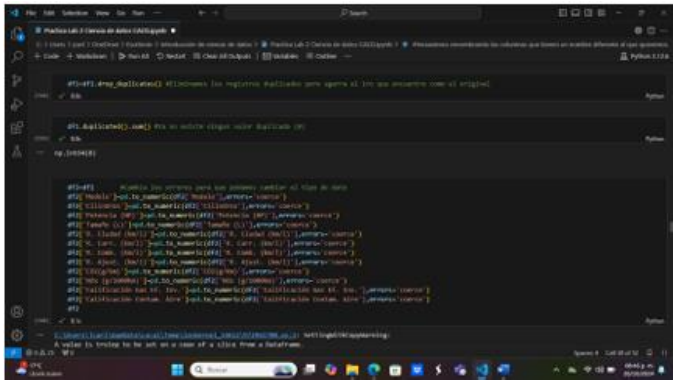
Conclusión

1. Como primeros cambios que note sería la correción del nombre de las columnas, ya que tienen caracteres que simplemente hacen que no se entienda de manera eficaz en que consiste cada columna.
2. Con la función "isnull().sum()" note que en todas las columnas existen datos que no tienen valores, por lo que sera necesario realizar algo para que no nos afecten en un futuro.
3. Con la función "duplicated()" encontramos que existen valores repetidos, por lo que sera necerio quitarlos.
4. Con la función "info()" podemos observar que la mayoría de las columnas los Dtype estan en object, lo cual en unas si es necesario pero en otras donde lo que mas se presentan son números, nos sería de mucha ayuda el que esten en tipo "int" o "float" de acuerdo al caso.
5. Al realizar las gráficas para apreciar mas a detalle cuales son los datos que tenemos se puede notar que existe un dato "aaaaa" lo cual no nos dice nada por lo que lo mejor sería removerlo.
6. El primer error con las gráficas es que algunas no se pudieron graficar por el simple hecho de que son de tipo "str", por lo que fue necesario ocupar la función ".unique()", con esto nos dimos cuenta que si existe el dato "aaaaa" que tendremos que reemplazar además de que en unas columnas hay datos "nan" los cuales tendremos que realizar algo con ellos

7. Inicie renombrando las columnas que tenían un nombre que no correspondía, esto con el comando rename.



8. Con el comando “drop_duplicates ()” eliminamos los registros duplicados, para que con “duplicated (). sum ()” nos aparezca el que ya no existe ningún valor duplicado. Ahora, con la función “pd.to_numeric [], errors=’coerce’” cambiaremos los errores por algo “numérico” y así poder cambiar el tipo de dato, esto solo funciono en las columnas que sus datos eran de puros números, nada de texto.

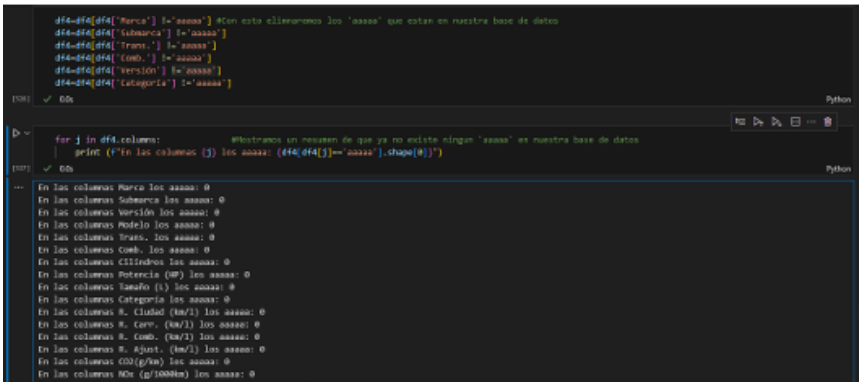


9. Ya que cambiamos nuestros datos en las columnas de datos numéricos a algo numérico valga la redundancia, utilice el comando “fillna (). mean (), inplace=True” para que sustituya los ‘nan’ por el promedio de cada columna a la cual escribamos.

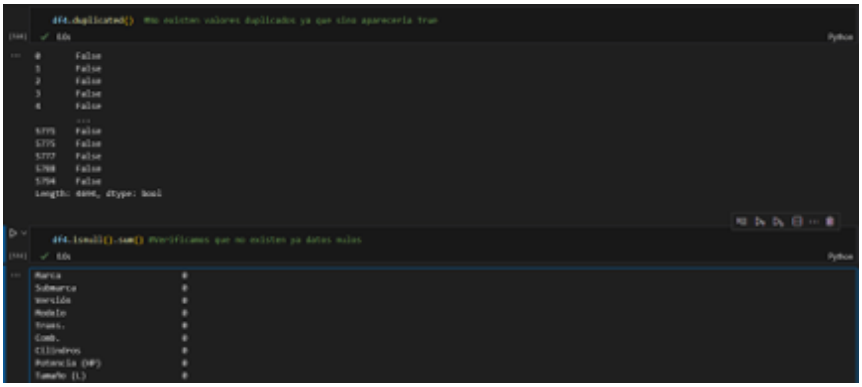
10. Ahora, con respecto a las columnas las cuales sus datos son textos, lo que utilice para eliminar los nan fue primero utilizar la función “fillna (0)” para cambiar los ‘nan’ por 0.

11. Utilizando la función “replace” cambiaremos los 0 que anteriormente habíamos puesto con el “fillna (0)” por la palabra ‘Sin_información’, esto para que en un futuro no nos genere confusión que entre vario texto existan 0.

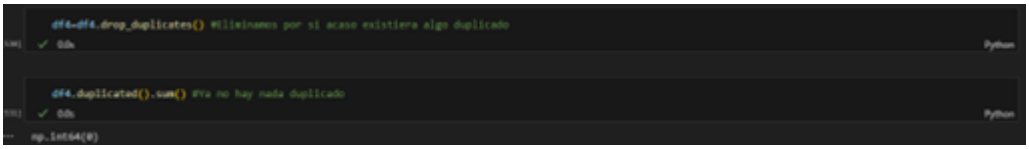
12. ¡Como recordemos anteriormente supimos de la existencia de un dato “aaaaa” por lo que decidí eliminarlo, para eso ocuparemos la función “! ='aaaaa’” para que elimine ese dato en cada columna, en este caso solo será en las de texto. Ya después con un para haremos un resumen para que ahora sí, ya no existe ese dato en nuestra base.



13. Ya ahora ocupamos los distintos comandos para asegurarnos que ya no existen valores “raros” en nuestra base de datos. “duplicated ()” para saber la existencia de valores duplicados, “isnull (). sum ()” para verificar que ya no existen datos nulos.

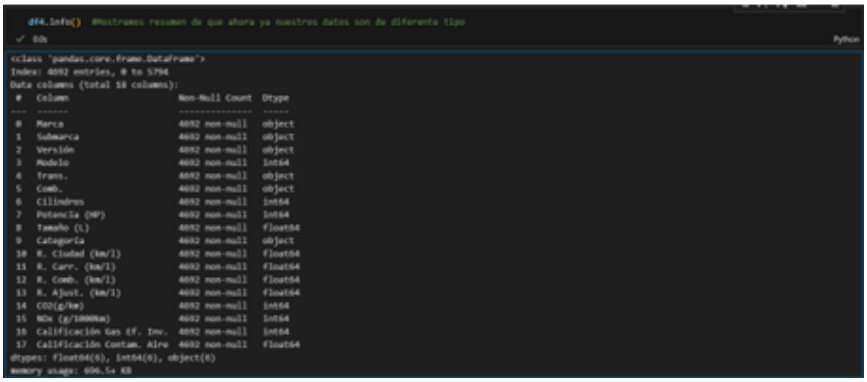


14. Ya por último realicé una búsqueda de filas duplicadas por si las dudas.

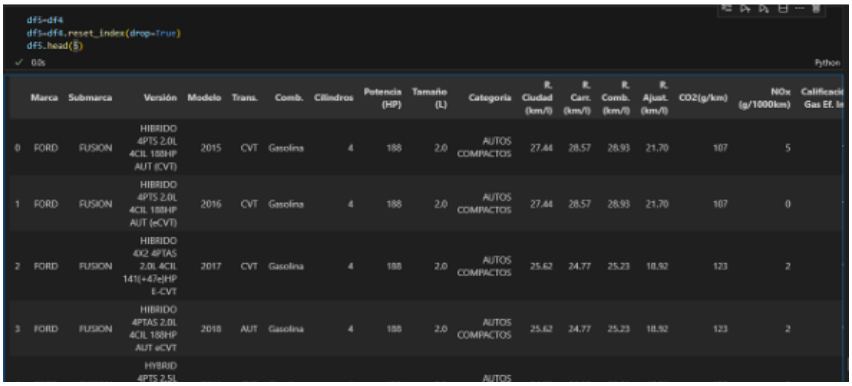


15. Ya que nuestra base quedo limpia, procedí a cambiar con ayuda del comando “. astype(int)” o “. astype(float)” de acuerdo al caso para cambiar el tipo de dato ya que antes la mayoría estaba en object, pero como recordamos la mayoría de columnas son de números.

16. Realizamos un resumen para mostrar que ya se han cambiado los tipos de datos. Esto con la función ".info ()".



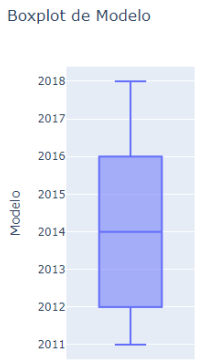
17. Ya que nuestra base ha quedado lista procedemos a reindexar el índice esto con el comando (. reset_index[drop=True]) esto para que no existan "saltos" entre números de fila.



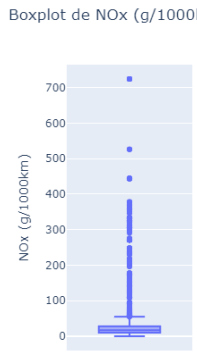
Manejo de Valores atípicos.

Para manejar estos valores lo que se realizado fue que se realizó la búsqueda de estos valores en cada columna con la ayuda de boxplots, esto para cada columna, en algunas si se encontraron y en otras no por lo que después de realizar los

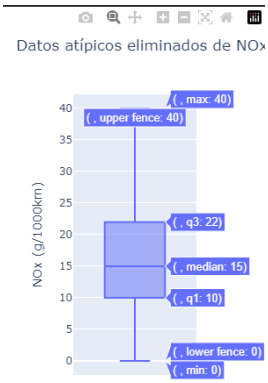
boxplots, en las columnas que se encontraban estos valores se corregían para así tener una base de datos limpia. Adelante se mostrará los ejemplos.



Como en este ejemplo, podemos apreciar que no existe ningún dato atípico por lo que no es necesario realizar alguna corrección.



Ahora tomemos de ejemplo esta columna, tiene muchísimos valores atípicos por lo que será necesario el corregirlo. Una vez corregido obtenemos el siguiente boxplot.



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 mas 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.

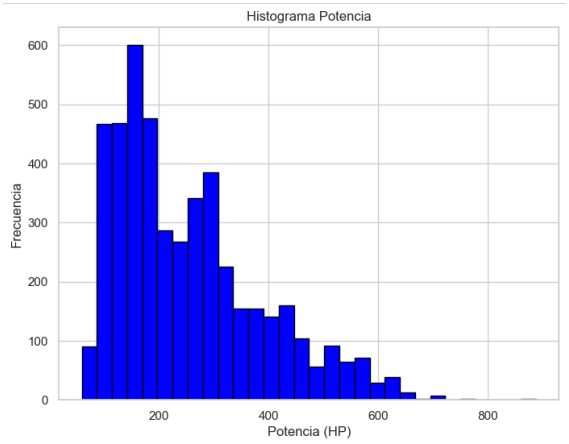
	Modelo	Cilindros	Potencia (HP)	Tamaño (L)	R. Ciudad (km/l)	R. Carr. (km/l)	R. Comb. (km/l)	R. Ajust. (km/l)	CO2(g/km)	NOx (g/1000km)	Calificación Gas Ef. Inv.	Calificación Contam. Aire
count	4692.000000	4692.000000	4692.000000	4692.000000	4692.000000	4692.000000	4692.000000	4692.000000	4692.000000	4692.000000	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
```

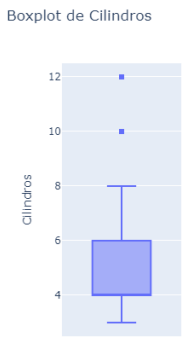
```
La mediana de la columna 'Potencia (HP)' es: 220.0
```

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 detecto 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.

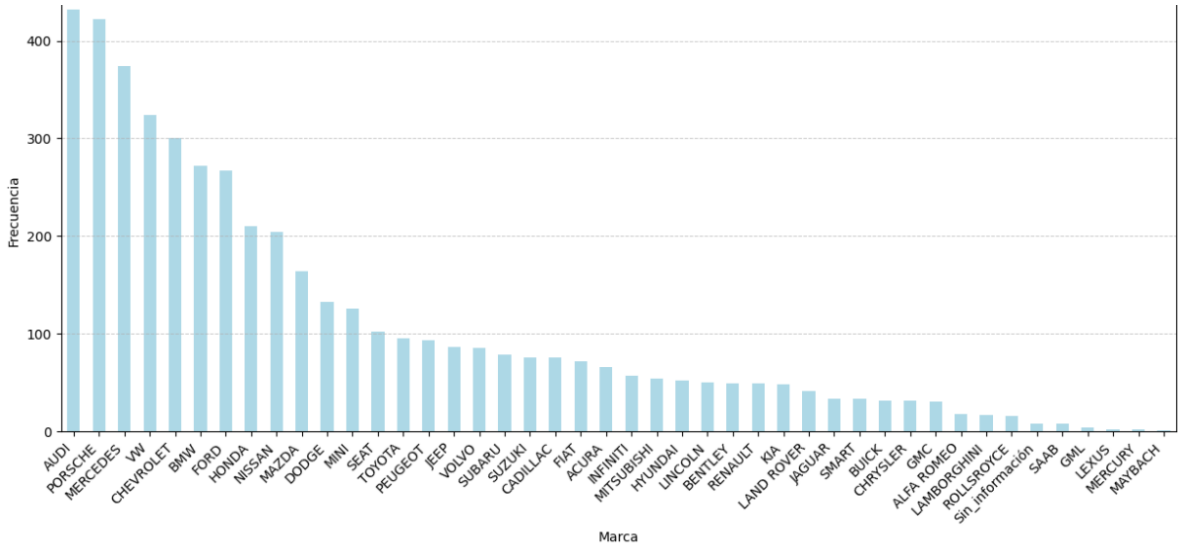


Ahora tenemos un histograma que para ser más específicos corresponde a la columna de Potencia (HP), al verla podemos notar que la mayoría de nuestros datos están en un rango de (30-650) aproximadamente, aunque podemos notar que existen unos ligeros datos después del 650, por lo que son datos que son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



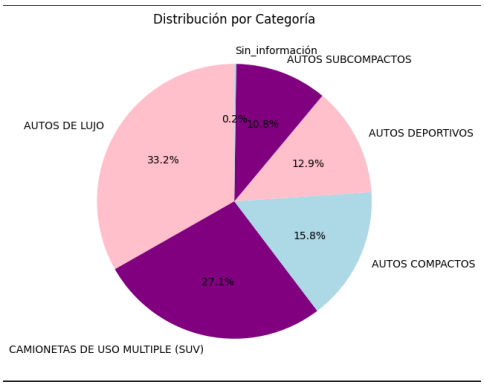
Ahora tenemos este Boxplot que pertenece a la columna cilindros, en ella rápidamente apreciamos 2 puntos en 12 y 10, esto principalmente a que en México como anteriormente se comento al sacar la mediana y promedio, los autos que circulan en su mayoría son de 4 cilindros, haciendo que los de 10 y 12 sean un lujo, pues principalmente se sabe que entre mas 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 mas prevalece en nuestro dataset, mientras que marcas mas 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.

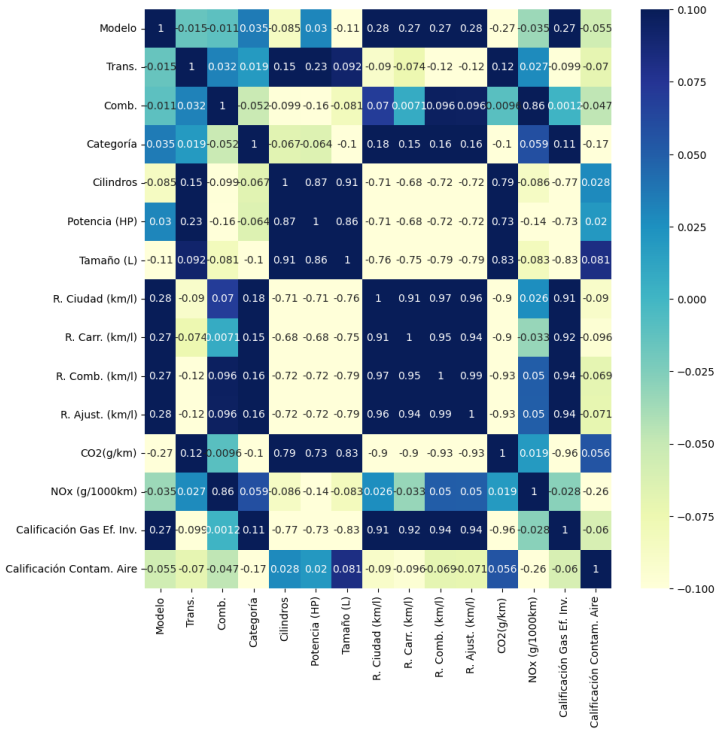


Ahora con esta grafica de pastel notamos que en México en la categoría de auto que mas hay en el país son la de autos de lujo, que bien, a pesar de que la palabra “Lujo” sea signo de caro, la mayoría de los autos son conocidos como lujo al incluir cosas que en otros no, aunque estas no sean tan costosas, 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.

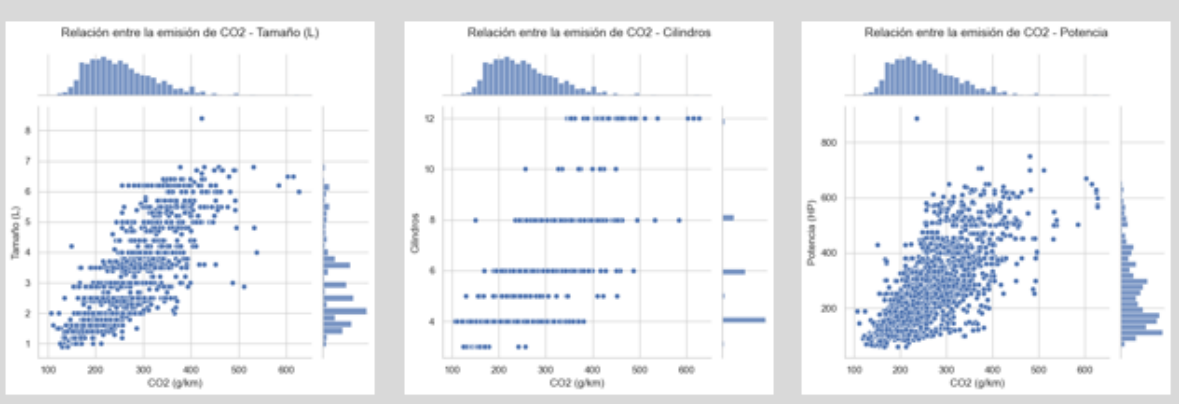
	Modelo	Trans.	Comb.	Categoría	Cilindros	Potencia (HP)	Tamaño (L)	Ciudad (km/l)	R. Carr. (km/l)	R. Comb. (km/l)	R. Ajust. (km/l)	CO2(g/km)	NOx (g/1000km)	Calificación Gas Ef. Inv.
Modelo	1.000000	-0.014930	-0.011140	0.035282	-0.085207	0.029787	-0.106898	0.275665	0.274931	0.273851	0.276575	-0.270076	-0.034924	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 mas 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 llego 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.



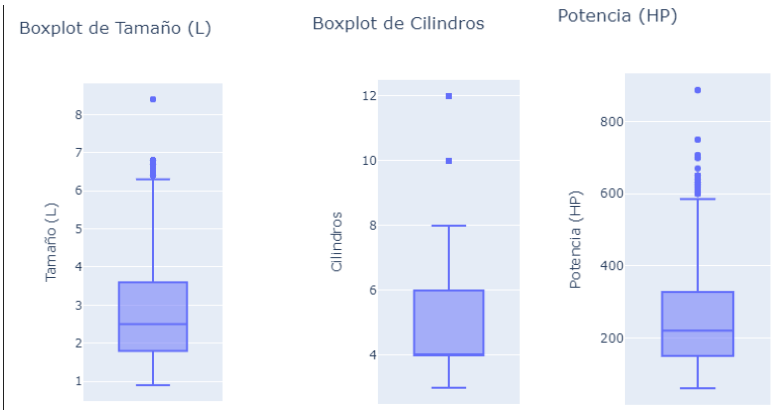
Ahora tenemos gráficos de dispersión, estos nos ayudan a entender de mejor forma como es que se relacionan las columnas ya mencionadas con el CO2, podemos observar que todas entre mayor sea tamaño, potencia o los cilindros mas generación de CO2 se da.

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,

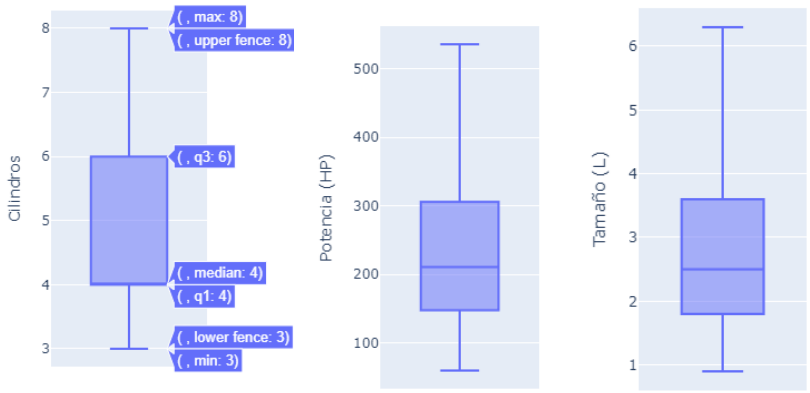


Tomemos como ejemplo este boxplot que pertenece a la columna tamaño, columna que se sabe que se relaciona con lo que buscamos que es el CO2.

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.

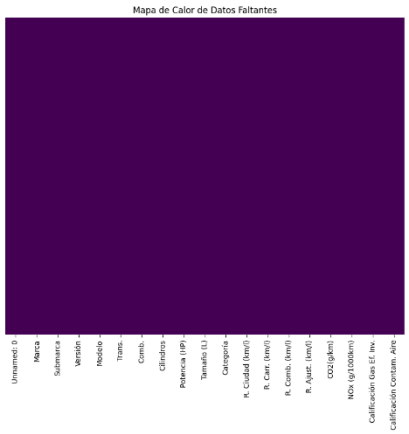


5. Análisis de valores faltantes

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.

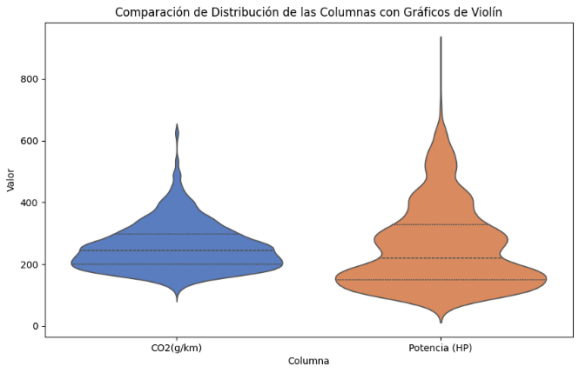
En el pasado algunos datos nulos o faltantes fueron eliminados y otros fueron cambiados a la palabra “Sin_información”.

```
# 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()
```



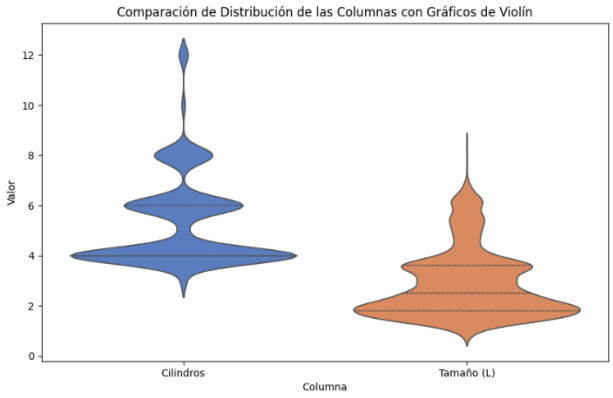
6. Relación entre variables categóricas y numéricas

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 mas 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 mas sea de algo, más subirá el CO2.

Modelo de Machine Learning

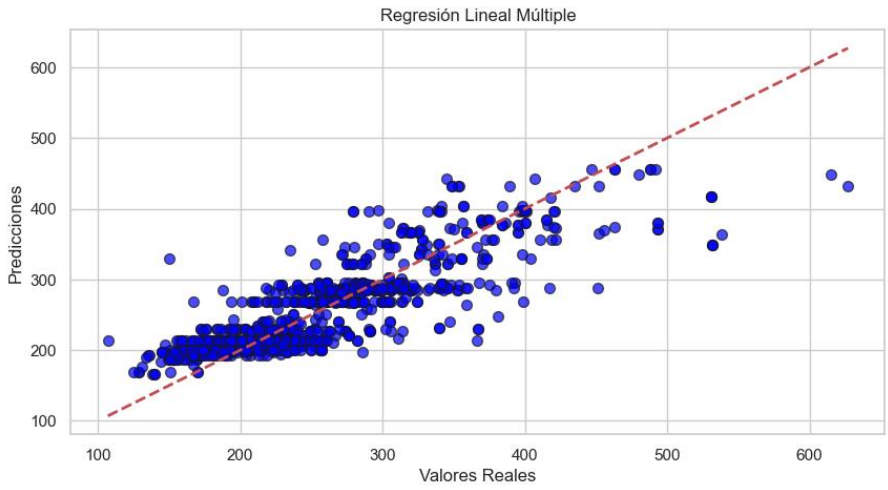
Descripción del modelo: El nombre del modelo elegido es Regresión lineal múltiple.

Justificación. Este tipo de modelo fue elegido ya que tenemos 3 variables que hacen que nuestra 4ta variable (CO2(g/km)) dependa de estas 3 anteriores, haciendo que dependiendo de los valores que obtengan las 3 variables, cambien de una forma a esta última variable.

Implementación y entrenamiento. Para saber con exactitud cuales eran los datos que se tenían que utilizar se realizaron varios cambios para así dividir los datos, principalmente el objetivo fue realizar un heatmap, pues este nos ayuda más rápido a saber que es lo que buscamos. Por lo que se calculo una matriz correlación con las columnas numéricas y esto hizo que nuestro mapa de calor nos dividiera los datos y su “importancia” a lo que buscamos.

De ahí se obtuvieron las 3 columnas mas importantes que se relacionan con el CO2.

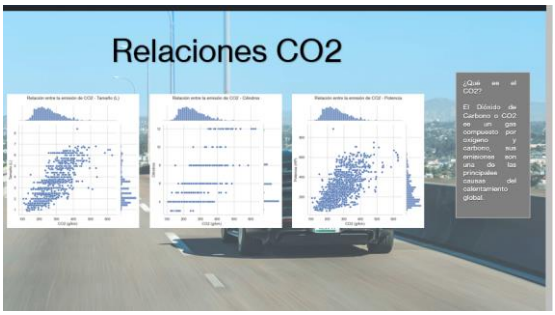
Resultados. Ya calculando nuestro modelo de regresión lineal múltiple se obtuvo que sigue siendo una línea diagonal de forma ascendente, es decir, parte de nuestras hipótesis eran ciertas.

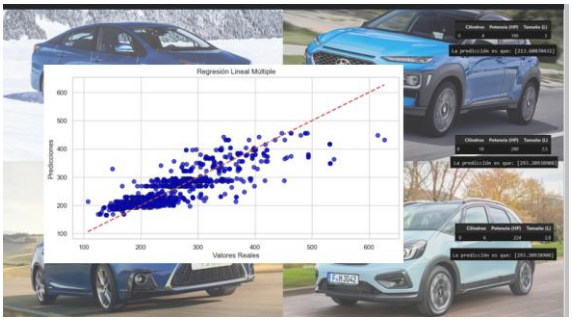


Dashboard

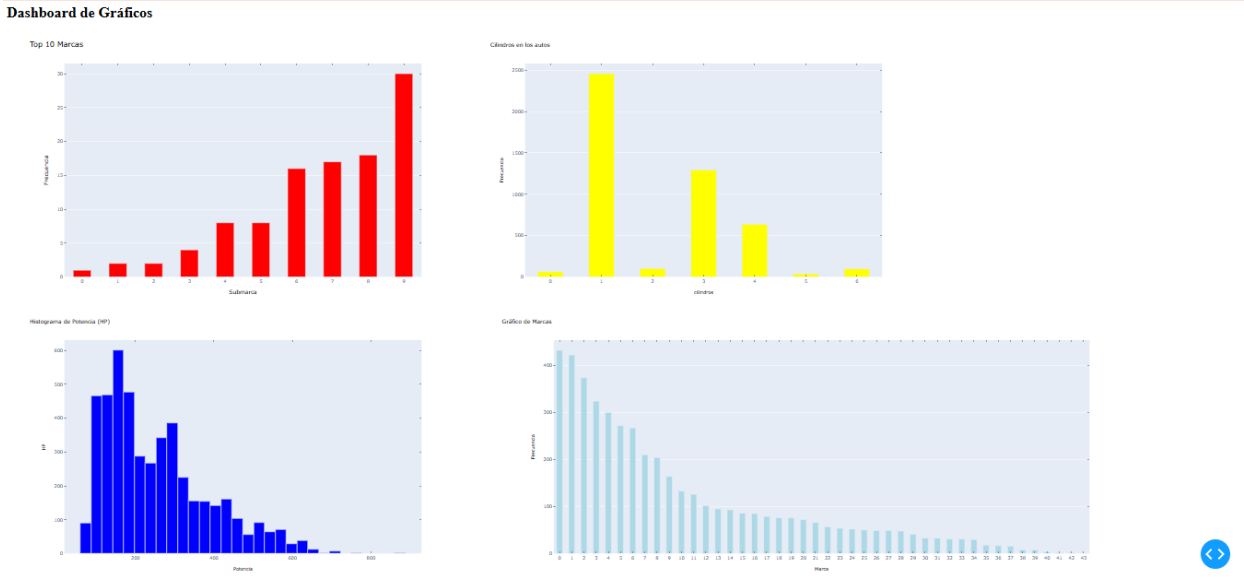
Explicación: El objetivo del dashboard es dar a los usuarios interesados una forma fácil y rápida de entender todo el objetivo al que se llegó después de la investigación y análisis de nuestra dataset. Se incluyeron 3 páginas ya que la primera corresponde a los datos en su forma general, la segunda los datos que se relacionan a lo que buscamos y finalmente en la tercer pagina el modelo de machine learning además de predicciones.

Uso y Beneficios: Estos datos están dirigidos hacia toda persona interesada en la contaminación que generan los autos, principalmente la generación del CO2, y opino que para todo publico pues desde pequeños podemos iniciar a explicar a las personas que es lo que en los autos provoca que se genere este gas, para así crear conciencia en la sociedad del porque los autos deben estar con un buen mantenimiento, ya que pedir que no se ocupen es algo imposible pues se sabe que hoy son una necesidad.





NOTA: Hubo un intento de dashboard que fue creado en Python, pero no me convenció, adjunto evidencias.



Conclusiones y futuras líneas de trabajo

Resumen de los hallazgos principales.

Después de una larga exploración los principales hallazgos fueron el como depende mucho los factores del automóvil para determinar su rendimiento, velocidad, etc. Pero nosotros principalmente nos fijamos en la generación del dióxido de carbono, donde después de todo lo ya explicado se llega a la conclusión de que entre mayor sea la potencia, tamaño y cilindros del automóvil, mayor generación de dióxido de carbono será.

Por lo que en parte si resolvió de manera general nuestras expectativas que teníamos al inicio, además de que varias hipótesis que fueron planteadas fueron comprobadas a lo largo del proyecto, aunque poco a poco se fueron aprendiendo mas cosas, como el como se relacionan las variables para hacer que una dependa de estas.

Posibles mejoras.

Tengo recomendaciones para los datos y es el agregar mas cosas del motor, podríamos agregar alguna comparación ya mas especifica entre marcas de refacciones de autos, donde podamos comparar que tan cierto es lo que nos dicen, si entre mejor sea la calidad, mejor será el rendimiento y por ende una menor contaminación.

También se podría investigar mas a fondo sobre las marcas y así comparar alguna marca que apenas este iniciando con una que lleve mas tiempo en el mercado, generando el si entre mas antigua la marca más experiencia habrá generado o simplemente las nuevas marcas se preocupan mas por el ambiente pues inician en una fecha donde la contaminación es un severo problema.

Referencias.

IBM (29 febrero 2024) “Regresión lineal múltiple” recuperado de https://www.ibm.com/docs/es/cognos-analytics/11.1.0?topic=tests-multiple-linear-regression el 23 de noviembre de 2024

IBM “¿Qué es el análisis exploratorio de datos (EDA)? Recuperado de https://www.ibm.com/mx-es/topics/exploratory-data-analysis el 23 noviembre de 2024

Acosta A (2017) “Autos – Consumo Gasolina México” recuperado de https://www.kaggle.com/datasets/checoalejandro/autos-consumo-gasolina-mexico el 20 de octubre de 2024

BBVA (06 octubre 2024) “¿Qué es el dióxido de carbono (CO2) y cómo impacta en el planeta?” recuperado de https://www.bbva.com/es/sostenibilidad/que-es-el-dioxido-de-carbono-co2-y-como-impacta-en-el-planeta/ el 23 de noviembre de 2024

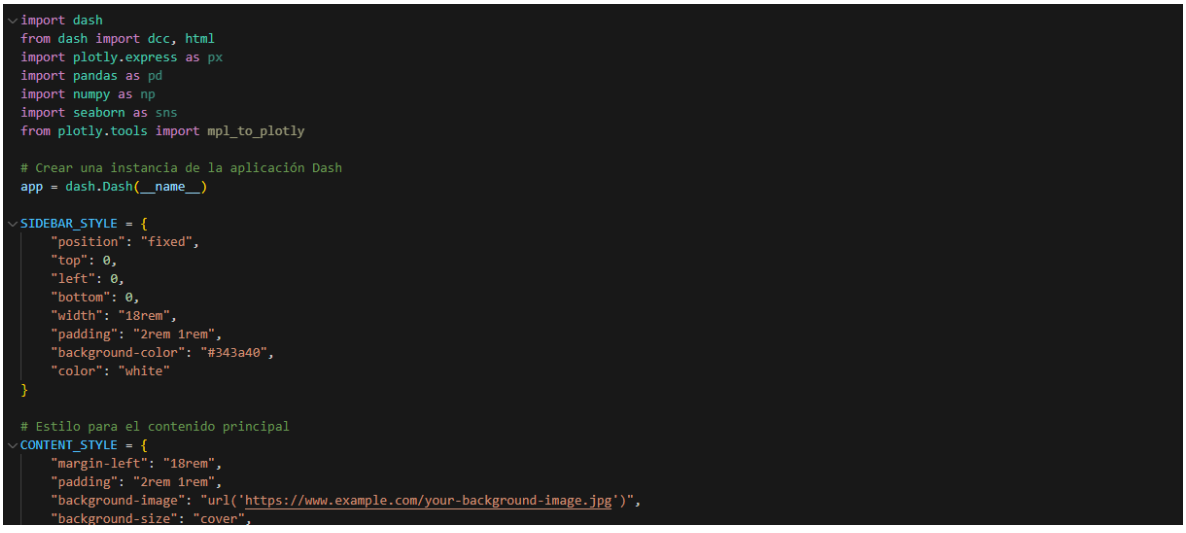
Link del dashboard

http://127.0.0.1:8050/

Anexos

Código fuente relevante para la implementación del modelo o limpieza de datos.

Código para generar el dashboard



```
import dash
from dash import dcc, html
import plotly.express as px
import pandas as pd
import numpy as np
import seaborn as sns
from plotly.tools import mpl_to_plotly

# Crear una instancia de la aplicación Dash
app = dash.Dash(__name__)

SIDEBAR_STYLE = {
    "position": "fixed",
    "top": 0,
    "left": 0,
    "bottom": 0,
    "width": "18rem",
    "padding": "2rem 1rem",
    "background-color": "#343a40",
    "color": "white"
}

# Estilo para el contenido principal
CONTENT_STYLE = {
    "margin-left": "18rem",
    "padding": "2rem 1rem",
    "background-image": "url('https://www.example.com/your-background-image.jpg')",
    "background-size": "cover",
```

```python
# Contenido principal
content = html.Div(id="page-content", style=CONTENT_STYLE)


# Definir tus gráficos aquí (reemplaza estos con tus gráficos reales)
# Ejemplo de tus gráficos ya generados (fig1, fig2, fig3, etc.)
# Ejemplo:
fig1c=fig1c
fig2c=fig2c
fig3c=fig3c
fig4c=fig4c
fig5c=fig5c
fig6c=fig6c



# Estructura del layout con tus gráficos
app.layout = html.Div(children=[
    html.H1("Dashboard de Gráficos"),

    # Primera fila con dos gráficos
    html.Div(children=[
        dcc.Graph(figure=fig1c),  # Aquí coloca tu gráfico fig1
        dcc.Graph(figure=fig4c),  # Aquí coloca tu gráfico fig2
    ], style={'display': 'flex', 'flex-wrap': 'wrap'}),

    # Segunda fila con dos gráficos
    html.Div(children=[
        dcc.Graph(figure=fig5c),  # Aquí coloca tu gráfico fig3
```

```python
    # Segunda fila con dos gráficos
    html.Div(children=[
        dcc.Graph(figure=fig5c),  # Aquí coloca tu gráfico fig3
        dcc.Graph(figure=fig6c),  # Aquí coloca tu gráfico fig4
    ], style={'display': 'flex', 'flex-wrap': 'wrap'}),
])

# Ejecutar el servidor de Dash
if __name__ == '__main__':
    app.run_server(debug=True)
```

Código para generar el modelo de machine learning

```python
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score

# Crear datos de ejemplo
X = df2[['Cilindros','Potencia (HP)','Tamaño (L)']]
# Generar una variable objetivo con algo de ruido
y = df2['CO2(g/km)']

# Dividir el conjunto de datos en entrenamiento y prueba
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)

# Crear y entrenar el modelo de regresión lineal múltiple
model = LinearRegression()
model.fit(X_train, y_train)

# Hacer predicciones
y_pred = model.predict(X_test)


# Evaluar el modelo
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)
print(f"Mean Squared Error: {mse:.2f}")
print(f"R^2 Score: {r2:.2f}")
```

```python
# Graficar resultados
plt.figure(figsize=(10, 5))
plt.scatter(y_test, y_pred, color="blue", edgecolor="k", s=50, alpha=0.7)
plt.plot([y.min(), y.max()], [y.min(), y.max()], 'r--', lw=2)
plt.xlabel("Valores Reales")
plt.ylabel("Predicciones")
plt.title("Regresión Lineal Múltiple")
plt.show()
```

Código para generar el heat map.

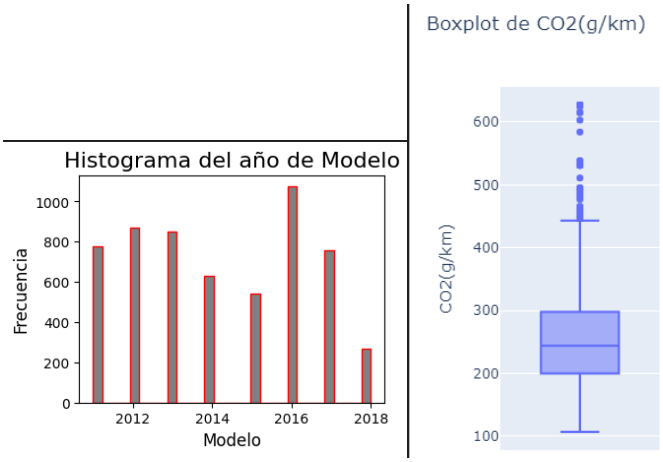
```
#Mapa de calor
plt.figure(figsize=(10,10))
#Grafico de calor con rango limitado
sns.heatmap(correlation_matrix2, annot=True, cmap='YlGnBu', vmin=-0.1, vmax=0.1)

# Mostrar gráfico
plt.show()
```

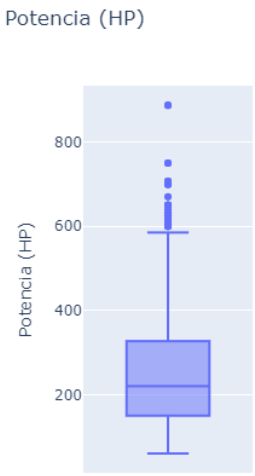
Código para generar la matriz correlación

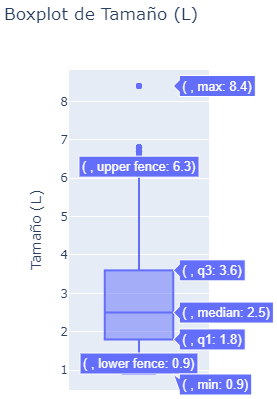
```
#Seleccionar solo las columnas númericas
df2_numeric = df2[['Modelo','Trans.', 'Comb.','Categoría','Cilindros','Potencia (HP)','Tamaño (L)','R. Ciudad (km/l)','R. Carr. (km/l)',
                   'R. Comb. (km/l)','R. Ajust. (km/l)','CO2(g/km)','NOx (g/1000km)','Calificación Gas Ef. Inv.','Calificación Contam. Aire']]
#Calcular la matriz correlación
correlation_matrix2 = df2_numeric.corr()
correlation_matrix2
```

Gráficos o análisis adicionales no incluidos en el reporte principal.

En estos gráficos se pueden apreciar un histograma de la columna “modelo”, además agregue unos boxplots con datos atípicos de las principales columnas que se tomaron en cuenta en nuestra base de datos.







Base de datos limpia que se utilizó

https://raw.githubusercontent.com/Carlos-Castillo-Gomez/Proyecto-Final-Ciencia-de-CACG-datos-/refs/heads/main/Base_limpia_proyecto_CACG

FIN