9hyhnqkgj

February 4, 2023

0.1 - Abstracto con Motivación y Audiencia

El presente análisis, surge de la necesidad de conocer los distintos tipos de combustibles que utilizan los automóviles actualmente y su grado de emisión de gases CO2, los cuales son contaminantes para nuestro medio ambiente. Los principales destinatarios de este informe, deberían ser los mismos fabricantes de automotores, así como también cualquier propietario de ellos y, las instituciones públicas relacionadas con el medio ambiente, ya que estas últimas deberían arbitrar y sancionar (de corresponder) a quienes fabriquen y posean automóviles con altos grados de emisión de CO2.

0.2 - Definición de Objetivo

Nuestro objetivo principal y final, es dar respuesta a la siguiente pregunta: ### ¿Cómo podemos reducir la emisión de gases contaminantes CO2 producidos por los automóviles?

0.3 - Contexto Comercial

De los gases que salen del caño de escape de cualquier auto con motor a combustión, hay uno que se ha convertido en el enemigo público número uno de la industria automotriz. Es el dióxido de carbono –CO2–, responsable de buena parte del Calentamiento Global que amenaza la vida humana tal cual la conocemos. La Unión Europea se puso a la vanguardia para combatirlo, con unas normativas extremadamente severas que incluyen multas multimillonarias a las automotrices. Y esto está acelerando los cambios en una industria que ya no volverá a ser la misma. Va siendo hora que nos metamos en un tema que es bastante más complejo de lo que parece, y que tarde o temprano nos terminará impactando.

0.4 - Problema Comercial

Las preguntas a responder, con respecto a este estudio, deberían ser por lo menos las siguientes:

. ¿Qué tipos de combustibles se utilizan actualmente en los automóviles?

. ¿Qué alternativas existen, hoy en día, a los motores que funcionan por combustión?

. ¿Cuales son, principalmente, los combustibles que generan mayores niveles de CO2?

. ¿A mayor cilindrada y tamaño de motor, es mayor el grado de emisión de CO2?

. ¿Cuál es el nivel de variación de emisión de CO2 que se dió en los últimos años?

. Los fabricantes de automóviles, ¿han tenido consideración realmente de este aspecto, a lo largo de los años?

. ¿Cuál fue la evolución de las multas/sanciones impuestas por los Organismos de Control en esta materia?

. ¿Qué otras alternativas de movilidad/transporte encontramos actualmente? ¿Cuál ha sido su variación en el uso, durante la última época?

0.5 - Contexto Analítico

Para este trabajo, se seleccionó una base de datos que contiene registros de emisiones de dióxido de carbono de automóviles con datos de motores, combustibles, transmisiones, consumos, etc. Esta base, tiene registros con los datos de los automóviles comercializados durante 7 años (7685 registros). Al existir automóviles mecanicamente iguales, que se venden con diferente equipamiento, hay registros duplicados. Las emisiones de CO2 se miden en gr/km y en consumo combinado (ciudad y ruta).

0.6 - E.D.A.

Observamos la estructura y las primeras filas de nuestro Dataset:

[ ]:
```
CO2 = pd.read_excel("/content/CO2 Emissions (1).xlsx")
CO2.head(10)
```

[ ]:

	Make	Model	Vehicle Class	Engine_Size_L	Cylinders	Transmission	\
0	ACURA	ILX	COMPACT	2.0	4	AS5	
1	ACURA	ILX	COMPACT	2.4	4	M6	
2	ACURA	ILX HYBRID	COMPACT	1.5	4	AV7	
3	ACURA	MDX 4WD	SUV - SMALL	3.5	6	AS6	
4	ACURA	RDX AWD	SUV - SMALL	3.5	6	AS6	
5	ACURA	RLX	MID-SIZE	3.5	6	AS6	
6	ACURA	TL	MID-SIZE	3.5	6	AS6	
7	ACURA	TL AWD	MID-SIZE	3.7	6	AS6	
8	ACURA	TL AWD	MID-SIZE	3.7	6	M6	
9	ACURA	TSX	COMPACT	2.4	4	AS5	

	Fuel Type	FuelType1	Fuel Consumption City (L/100 km)	\
0	Z	Premium gasoline	9.9	
1	Z	Premium gasoline	11.2	
2	Z	Premium gasoline	6.0	
3	Z	Premium gasoline	12.7	
4	Z	Premium gasoline	12.1	
5	Z	Premium gasoline	11.9	
6	Z	Premium gasoline	11.8	
7	Z	Premium gasoline	12.8	
8	Z	Premium gasoline	13.4	
9	Z	Premium gasoline	10.6	

	Fuel Consumption Hwy (L/100 km)	Fuel Consumption Comb (L/100 km)	\
0	6.7	8.5	
1	7.7	9.6	

2	5.8	5.9
3	9.1	11.1
4	8.7	10.6
5	7.7	10.0
6	8.1	10.1
7	9.0	11.1
8	9.5	11.6
9	7.5	9.2

	Fuel Consumption Comb (mpg)	Fuel Consumption Comb (Kmpl)	Emissions
0	33	14.029752	196
1	29	12.329176	221
2	48	20.406912	136
3	25	10.628600	255
4	27	11.478888	244
5	28	11.904032	230
6	28	11.904032	232
7	25	10.628600	255
8	24	10.203456	267
9	31	13.179464	212

[ ]:
```
# Tamaño de nuestro Dataset: (7385 registros y 14 columnas/variables)
CO2.shape
```

[ ]: (7385, 14)

[ ]:
```
# Verifico datos duplicados debido a versiones del mismo automóvil con␣
  ↪diferente equipamiento.
CO2.duplicated().sum()
print('Hay un total de ' + (str(CO2.duplicated().sum()) + ' duplicados en el␣
  ↪dataset.'))
```

Hay un total de 1103 duplicados en el dataset.

[ ]:
```
# Borro entradas duplicadas
```

[ ]:
```
# La base de datos queda ahora de un tamaño menor:
CO2.shape
```

[ ]: (6282, 14)

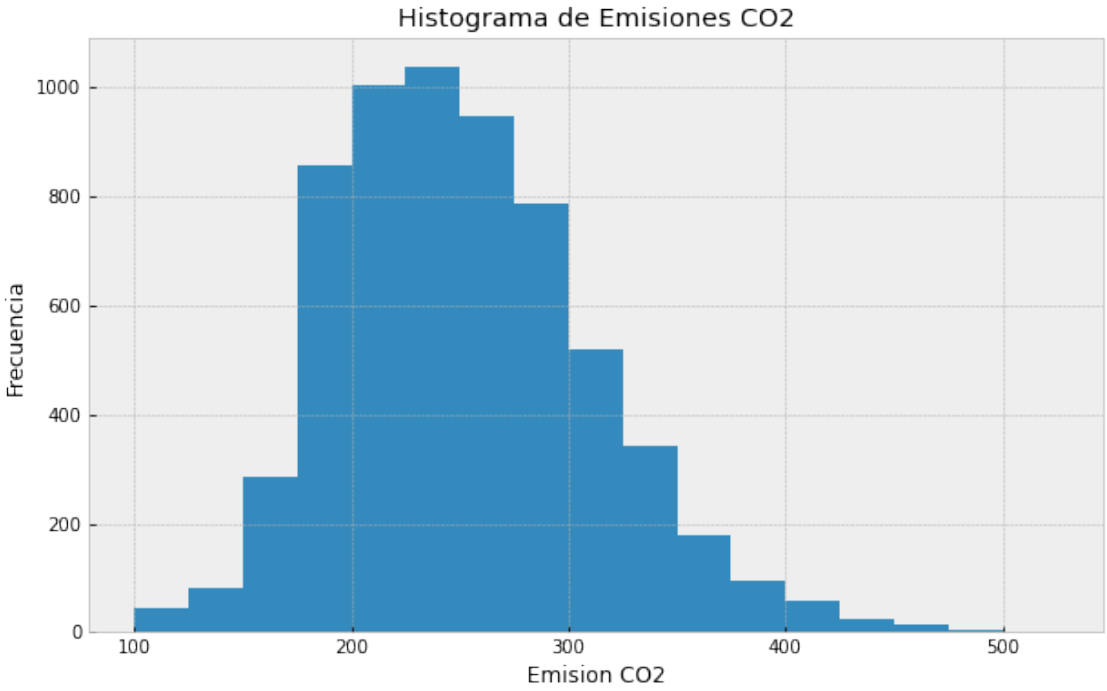
Se utilizarán los datos de la base para encontrar correlaciones entre distintas características de los automóviles y las emisiones de CO2.

Para iniciar, se buscará una correlación entre la cilindrada y el número de cilindros versus la emisión de CO2.

###Conociendo a la variable objetivo: Emissions - Emisiones de CO2:###

[ ]:
```
fig, ax = plt.subplots(figsize=(10,6))
ax.hist(x= CO2['Emissions'], bins=range(100,550,25), density=False,)
ax.set_xlabel('Emision CO2')
ax.set_ylabel('Frecuencia')
ax.set_title('Histograma de Emisiones CO2')
```

[ ]:
```
Text(0.5, 1.0, 'Histograma de Emisiones CO2')
```



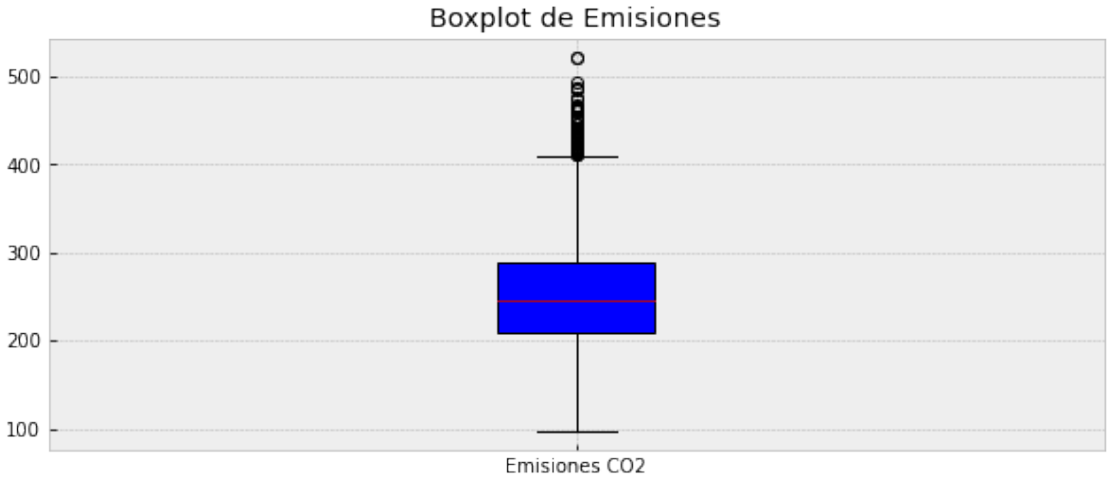
Se ve una distribución de datos del tipo gaussiana (asimétrica derecha) para la variable objetivo, con una media aproximada de 250.

[ ]:
```
# Se comprueba que la variable Emisiones de CO2 tiene una media de 250.

CO2.Emissions.mean()
```

[ ]:
```
251.1577523081821
```

[ ]:
```
# Chequeando cómo se distribuyen los valores para esta variable:

fig, ax = plt.subplots(figsize=(10, 4))
ax.boxplot(CO2['Emissions'], labels=["Emisiones CO2"],patch_artist=True)
ax.set_title('Boxplot de Emisiones')
```

[ ]:
```
Text(0.5, 1.0, 'Boxplot de Emisiones')
```



En este boxplot para la variable objetivo, Emisiones, se ven los cuartiles definidos y la mediana en 250 y los outliers para valores superiones a 400.

###Análisis Bivariado###

[ ]:
```
# Análisis entre una variable categórica (Tipo de Combustible o Fuel Tipe) con
↪nuestra varible objetivo que es del tipo numérica (emisiones de CO2 o
↪Emissions)
CO2.groupby('FuelType1')['Emissions'].mean().sort_values(ascending=False)
```

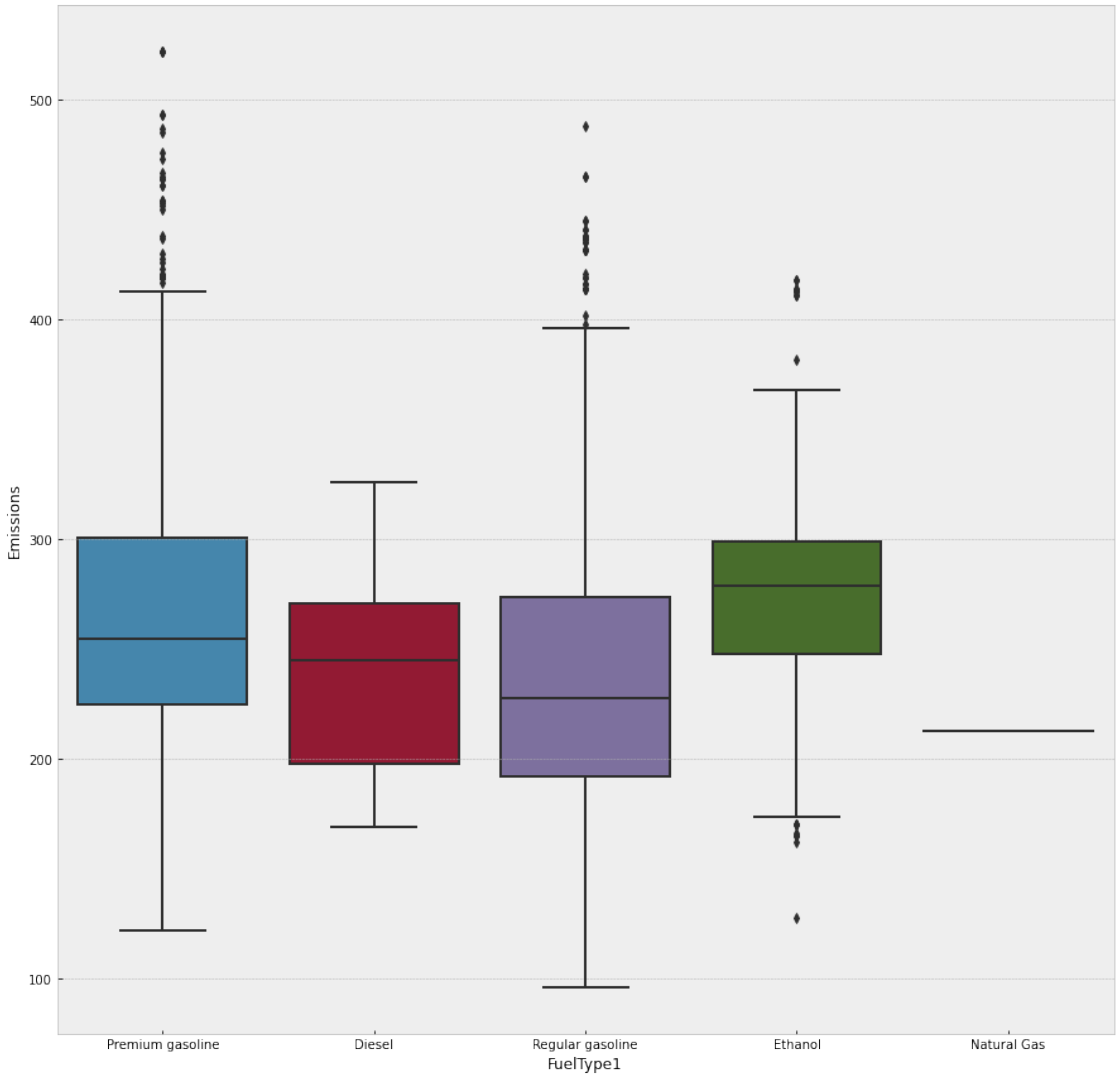
[ ]:
```
FuelType1
Ethanol             275.091892
Premium gasoline    266.043410
Diesel              237.548571
Regular gasoline    235.119329
Natural Gas         213.000000
Name: Emissions, dtype: float64
```

[ ]:
```
CO2.groupby('FuelType1')['Emissions'].describe()
```

[ ]:
```
                   count        mean        std    min    25%    50%    75%  \
FuelType1
Diesel             175.0  237.548571  41.817704  169.0  198.0  245.0  271.0
Ethanol            370.0  275.091892  47.093198  128.0  248.0  279.0  299.0
Natural Gas          1.0  213.000000        NaN  213.0  213.0  213.0  213.0
Premium gasoline  3202.0  266.043410  56.695972  122.0  225.0  255.0  301.0
Regular gasoline  3637.0  235.119329  57.401473   96.0  192.0  228.0  274.0

                   max
FuelType1
Diesel           326.0
```

```
Ethanol           418.0
Natural Gas       213.0
Premium gasoline  522.0
Regular gasoline  488.0
```

[ ]:
```
plt.figure(figsize=(15,15)) # defino el tamaño del grafico
sns.boxplot(y = 'Emissions', x = 'FuelType1', data = CO2)
plt.show()
```



La mediana tiene valores similares para cada tipo de combustible, excepto el gas natural, que hay un solo valor en el dataset.

0.6.1 Análisis de Correlaciones Lineales entre Variables Numéricas

```
[ ]: plt.figure(figsize=(12, 8))
     CO2 = CO2.corr()
     sns.heatmap(CO2,
                 xticklabels = CO2.columns.values,
                 yticklabels = CO2.columns.values,
                 annot = True);
```

Para la Variable Objetivo, Emisiones de CO2, vemos cómo se relaciona con el resto de las variables del dataset:

```
[ ]: CO2.corr()['Emissions']
```

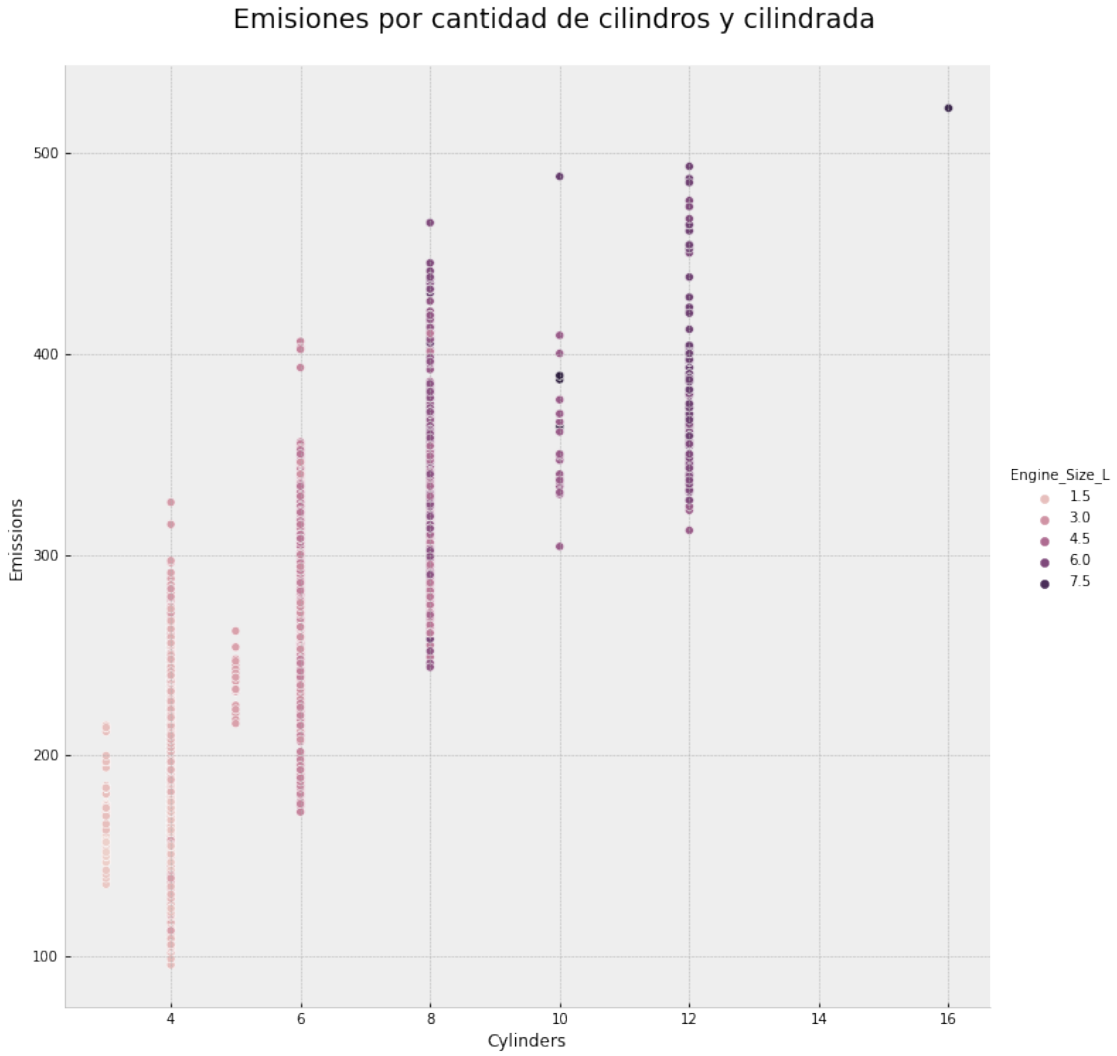
```
[ ]: Engine_Size_L                    0.990889
     Cylinders                       0.987259
     Fuel Consumption City (L/100 km)    0.997748
     Fuel Consumption Hwy (L/100 km)     0.994656
     Fuel Consumption Comb (L/100 km)    0.997014
     Fuel Consumption Comb (mpg)        -0.998334
     Fuel Consumption Comb (Kmpl)       -0.998334
```

```
Emissions                       1.000000
Name: Emissions, dtype: float64
```

###Análisis Multivarido###

Empecemos a ver la relación de nuestra variable objetivo, emisiones de CO2 ("Emissions"), con otras del dataset:

[ ]:
```
# Análisis de correlación entre la cantidad de cilindros y la emisión, con detalle en la cilindrada
cilindrada_co2 = sns.pairplot(
    CO2,
    x_vars = ['Cylinders'],
    y_vars = ['Emissions'],
    hue = 'Engine_Size_L',
    height = 10
)
cilindrada_co2.fig.suptitle("Emisiones por cantidad de cilindros y cilindrada", y = 1.04, fontsize = 19);
```



Se puede observar que existe una esperada correlación entre la cilindrada y el número de cilindros, con la emisión de dióxido de carbono.

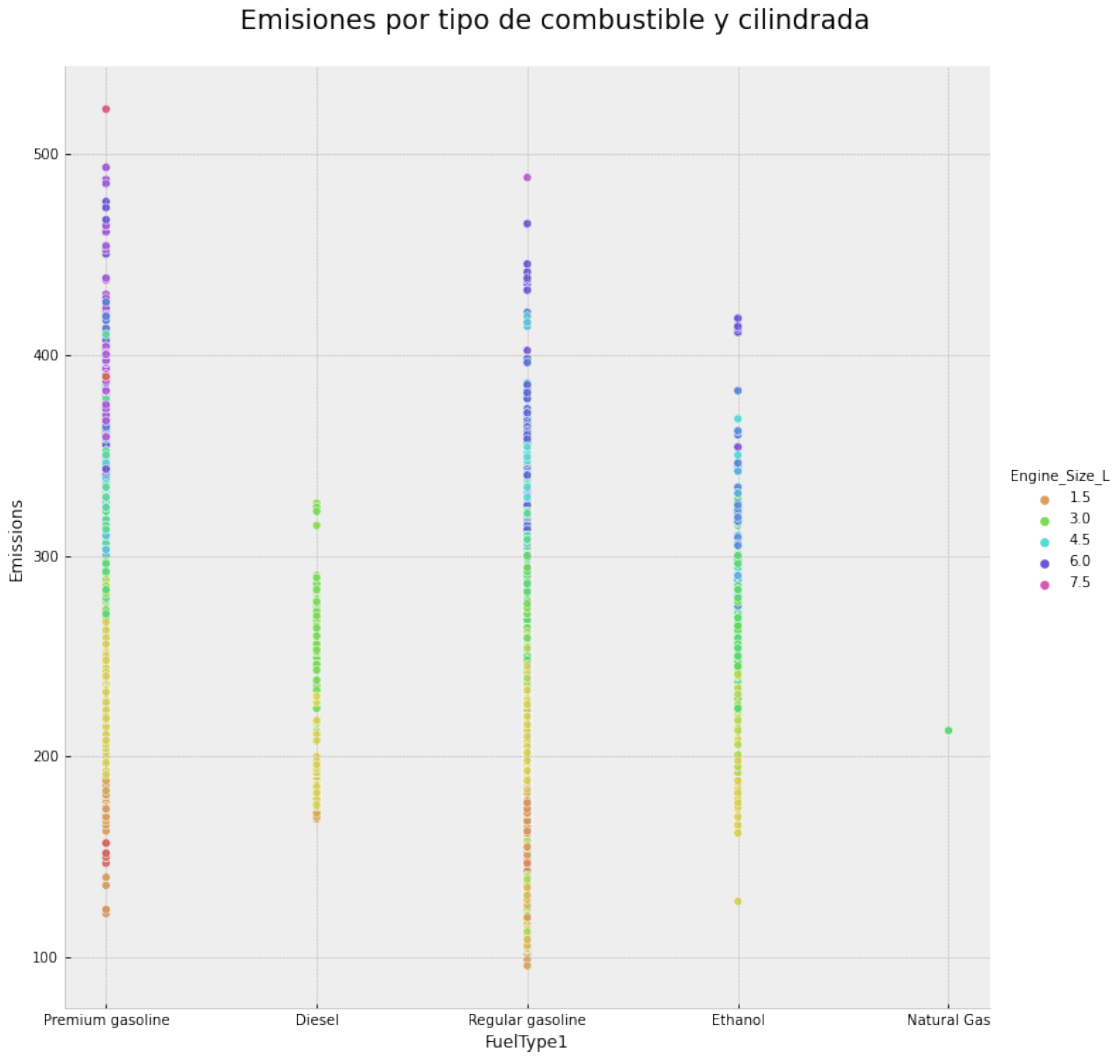
Motores más grandes y con mayor número de cilindros generan más emisiones.

Los motores más eficientes son los de 4 cilindros con cilindradas entre 1.4 y 2 litros.

En el segundo análisis, se buscará una posible correlación entre el tipo de combustible y la emisión, para determinar qué combustible es el que genera menor emisión.

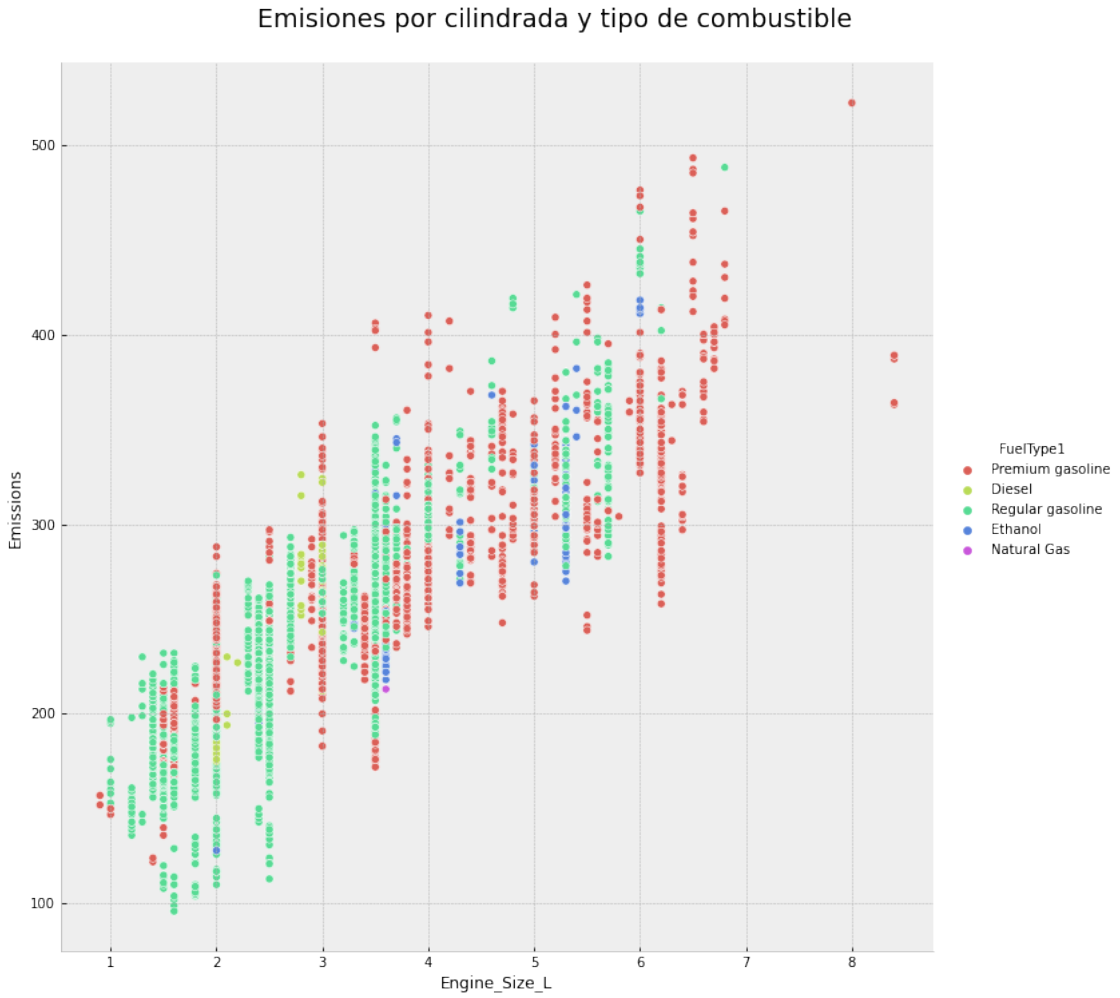
[ ]:
```
# Análisis de correlación entre el tipo de combustible y la emisión, con␣
 ↪detalle en la cilindrada
combustible_co2 = sns.pairplot(
    CO2,
    x_vars = ['FuelType1'],
    y_vars = ['Emissions'],
```

```
    hue = 'Engine_Size_L',
    palette="hls",
    height = 10
)
combustible_co2.fig.suptitle("Emisiones por tipo de combustible y cilindrada",␣
 ↪y = 1.04, fontsize = 19);
```



[ ]:
```
# Analisis de correlacion entre la cilindrada y la emisión, con detalle en el␣
 ↪tipo de combustible:
cilindrada_co2 = sns.pairplot(
    CO2,
    x_vars = ['Engine_Size_L'],
    y_vars = ['Emissions'],
    hue = 'FuelType1',
```

```
    palette="hls",
    height = 10
)
cilindrada_co2.fig.suptitle("Emisiones por cilindrada y tipo de combustible", y␣
 ↪= 1.04, fontsize = 19);
```



Se observa que el combustible que genera menor emisión de CO2 por km es la nafta súper.

Nota: no se tienen suficientes datos de GNC como para sacar conclusiones con respecto a esta variante de combustible.

0.7 Procesamiento y Selección de Variables

Seleccionaremos, en una primera instancia, las variables numéricas (nuestra variable objetivo, es también de este tipo).

[ ]:
```
CO2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7385 entries, 0 to 7384
Data columns (total 14 columns):
 #   Column                            Non-Null Count  Dtype
---  ------                            --------------  -----
 0   Make                              7385 non-null   object
 1   Model                             7385 non-null   object
 2   Vehicle Class                     7385 non-null   object
 3   Engine_Size_L                     7385 non-null   float64
 4   Cylinders                         7385 non-null   int64
 5   Transmission                      7385 non-null   object
 6   Fuel Type                         7385 non-null   object
 7   Fuel Type1                        7385 non-null   object
 8   Fuel Consumption City (L/100 km)  7385 non-null   float64
 9   Fuel Consumption Hwy (L/100 km)   7385 non-null   float64
 10  Fuel Consumption Comb (L/100 km)  7385 non-null   float64
 11  Fuel Consumption Comb (mpg)       7385 non-null   int64
 12  Fuel Consumption Comb (Kmpl)      7385 non-null   float64
 13  Emissions                         7385 non-null   int64
dtypes: float64(5), int64(3), object(6)
memory usage: 807.9+ KB
```

[ ]:
```
CO2_num = CO2.select_dtypes('number')
CO2_num.head()
```

[ ]:

	Engine_Size_L	Cylinders	Fuel Consumption City (L/100 km)	\
0	2.000	4	9.900	
1	2.400	4	11.200	
2	1.500	4	6.000	
3	3.500	6	12.700	
4	3.500	6	12.100	

	Fuel Consumption Hwy (L/100 km)	Fuel Consumption Comb (L/100 km)	\
0	6.700	8.500	
1	7.700	9.600	
2	5.800	5.900	
3	9.100	11.100	
4	8.700	10.600	

	Fuel Consumption Comb (mpg)	Fuel Consumption Comb (Kmpl)	Emissions
0	33	14.030	196
1	29	12.329	221
2	48	20.407	136
3	25	10.629	255
4	27	11.479	244

0.7.1 Análisis y Procesamiento de la Variable Cylinders

[ ]:
```
# Cantidad de Valores Únicos que tiene
CO2_num.Cylinders.nunique()
```

[ ]:
```
8
```

[ ]:
```
# Valores Únicos que tiene
CO2_num.Cylinders.unique()
```

[ ]:
```
array([ 4,  6, 12,  8, 10,  5, 16,  3])
```

[ ]:
```
# Conteo de Registros por cada Valor
CO2_num.Cylinders.value_counts()
```

[ ]:
```
4     1973
6     1448
8      742
12      94
3       57
10      24
5       21
16       1
Name: Cylinders, dtype: int64
```

0.7.2 Transformaciones de Columnas

Como venimos mencionando a lo largo de nuestro análsis, vamos a tomar como variable objetivo a "Emissions". Para ello, a su vez, vamos a transfromarla y dividirla en dos segmentos:

0 = aceptable = valores menores a 200 g/km

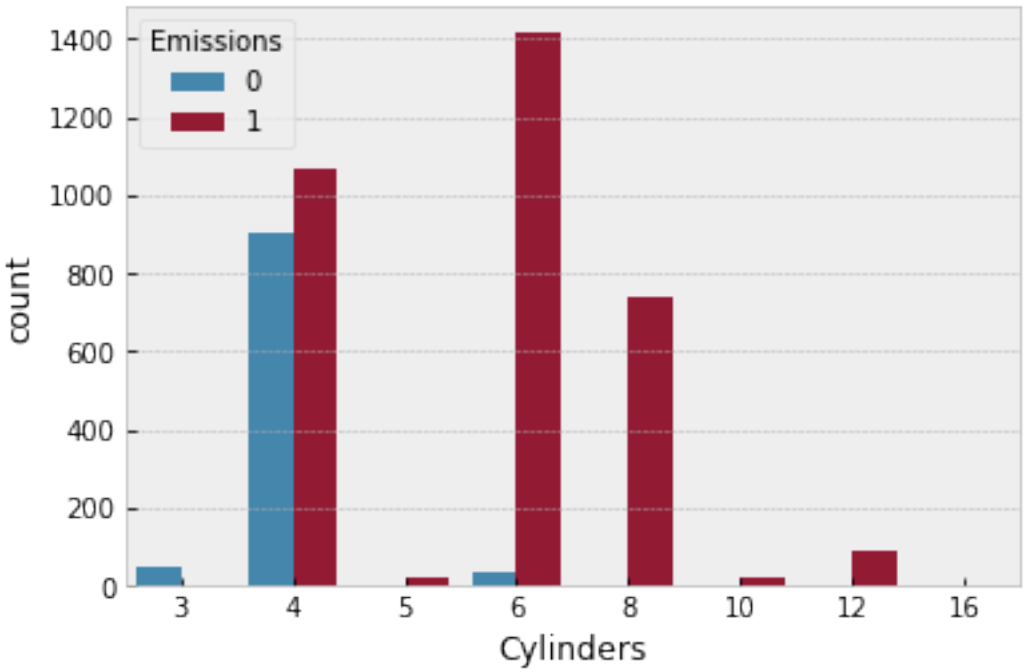
1 = no aceptable = valores mayores a 200 g/km

[ ]:
```
CO2_numcp.Emissions.value_counts()
```

[ ]:
```
1    3370
0     990
Name: Emissions, dtype: int64
```

[ ]:
```
sns.countplot(x='Cylinders', data = CO2_numcp, hue = 'Emissions')
```

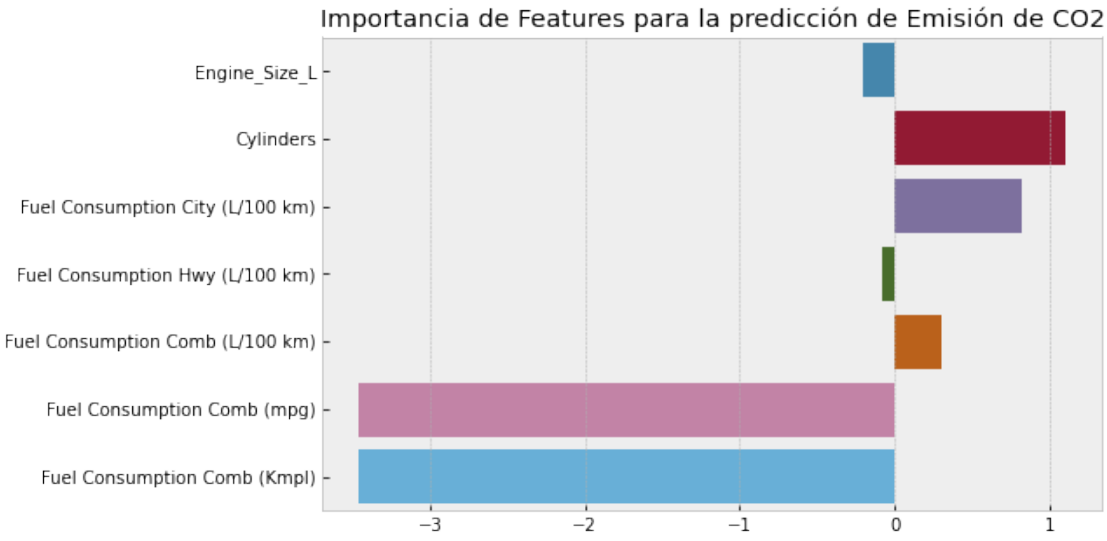
[ ]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f4ddbadf5d0>
```



Como se observa, claramente, el valor aceptable de nuestra variable objetivo se da práctimente sólo en los vehículos que poseen 6 o menos número de Cilindros.

0.8 Regresión Logística

[ ]:
```
plt.figure(figsize=(8,5))
values = pd.Series(clf.coef_.flat)
sns.barplot(y=X1.columns,
            x=values,
           ).set(title='Importancia de Features para la predicción de Emisión␣
 ↪de CO2')
plt.show()
```

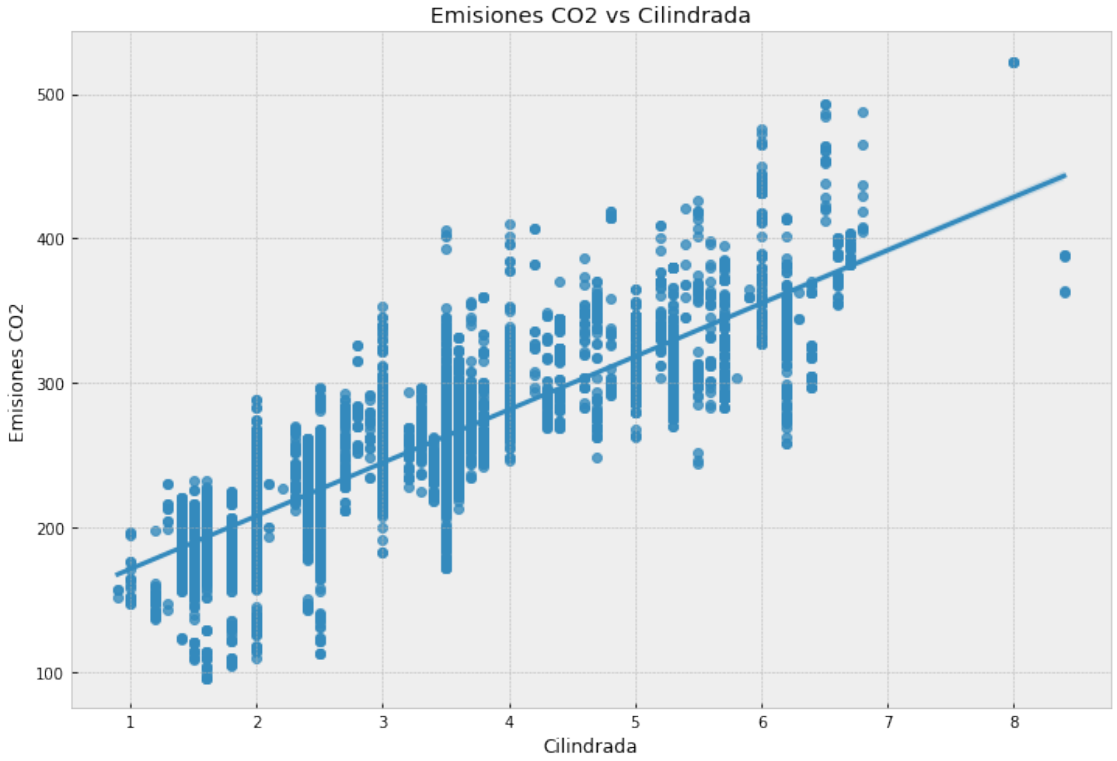


Con este último modelo, cambia drásticamente la importancia de las variables analizadas para explicar el comportamiento de la variable objetivo (acrecenta su importancia la variable "consumo combinado de combustible" tanto en mpg como en Kmpl, mientras que mantiene su relevancia en el análisis la variable "Cylinders", la cual hemos tratado en puntos anteriores).

Para la Variable Objetivo, Emisiones de CO2, vemos cómo se relaciona con el resto de las variables del dataset:

[ ]:
```
# Vemos cómo es el modelo de regresión lineal entre dos variables, nuestra␣
↪variable objetivo y el tamaño en litros del motor:
# Gráfico
ax = sns.regplot(x="Engine_Size_L", y="Emissions", data=CO2)
ax.set(xlabel='Cilindrada', ylabel='Emisiones CO2')
plt.title('Emisiones CO2 vs Cilindrada')
```

[ ]:
```
Text(0.5, 1.0, 'Emisiones CO2 vs Cilindrada')
```



[ ]:
```
# Ahora observamos los distintos indicadores resultantes de esta relación/
  ↪regresión:
model = smf.ols('Emissions ~ Engine_Size_L', data=CO2)
model = model.fit()
print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:              Emissions   R-squared:                       0.724
Model:                            OLS   Adj. R-squared:                  0.724
Method:                 Least Squares   F-statistic:                 1.941e+04
Date:                Sat, 05 Nov 2022   Prob (F-statistic):               0.00
Time:                        19:41:12   Log-Likelihood:                -35770.
No. Observations:                7385   AIC:                         7.154e+04
Df Residuals:                    7383   BIC:                         7.156e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
=
                    coef    std err          t      P>|t|      [0.025
0.975]
------------------------------------------------------------------------------
-
```

```
Intercept       134.3659      0.908    148.056      0.000     132.587
136.145
Engine_Size_L    36.7773      0.264    139.321      0.000      36.260
37.295
==============================================================================
Omnibus:                      212.800   Durbin-Watson:                   0.895
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              529.756
Skew:                           0.076   Prob(JB):                    9.22e-116
Kurtosis:                       4.303   Cond. No.                         9.36
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
```

R2=0.724 quiere decir que la emisión de gases CO2 al ambiente se describe en un 72.4% de manera lineal en función del tamaño en litros del motor. p-value=0, entonces el intercepto no es cero. La prueba ómnibus dice que la varianza de los datos es bastante amplia, ya que el resultado es muy distinto a cero.