MODUL PRAKTIKUM

IS411 – DATA MODELING
PROGRAM SARJANA S1 SISTEM INFORMASI
FAKULTAS TEKNIK DAN INFORMATIKA

PROGRAM STUDI SISTEM INFORMASI
FAKULTAS TEKNIK DAN INFORMATIKA
UNIVERSITAS MULTIMEDIA NUSANTARA

Gedung B Lantai 5, Kampus UMN
Jl. Scientia Boulevard, Gading Serpong, Tangerang, Banten-15811 Indonesia
Telp: +62-21.5422.0808 (ext. 1803), email: ict.lab@umn.ac.id, web: umn.ac.id

MODUL 13
LINEAR REGRESSION

DESKRIPSI TEMA

1. Regression
2. Linear Regression
3. Least Squares Fit concept

CAPAIAN PEMBELAJARAN MINGGUAN (SUB-CAPAIAN PEMBELAJARAN)

Students are able to illustrate data modeling problems using classification techniques (C_4)

PENUNJANG PRAKTIKUM

1. Anaconda Navigator
2. Jupyter Notebook

(+ Perlengkapan Apd/Alat Pelindung Diri Yang Harus Digunakan, Jika Ada)

LANGKAH-LANGKAH PRAKTIKUM

Import Library

1. Lakukan langkah yang sama seperti minggu sebelumnya untuk meng-import library Numpy dan Pandas, pylab, matplotlib.pyplot.

Import packages yang diperlukan

```
In [1]:
1 import matplotlib.pyplot as plt
2 import pandas as pd
3 import pylab as pl
4 import numpy as np
5 %matplotlib inline
```

Import Data

2. Gunakan dataset **FuelConsumptionCo2.csv** dan masukkan ke dalam dataframe menggunakan Pandas. Berikan nama **df** .
3. Tampilkan informasi dataset dan isi data.

Menampilkan Statistik Data

Dataset FuelConsumption.csv yang diunduh mengandung model spesifik untuk rating konsumsi bahan bakar (fuel consumption) dan estimasi emisi karbon dioksida untuk kendaraan ringan yang dijual di Kanada.

4. Membaca data :

```
In [3]:
df = pd.read_csv("FuelConsumption.csv")

# melihat dataset
df.head()
```

Out[3]:

	MODELYEAR	MAKE	MODEL	VEHICLECLASS	ENGINESIZE	CYLINDERS	TRANSMISSION	FUELTYPE	FUELCONSUMPTION_CITY	FUELCONSUMPTION_HW
0	2014	ACURA	ILX	COMPACT	2.0	4	AS5	Z	9.9	6.
1	2014	ACURA	ILX	COMPACT	2.4	4	M6	Z	11.2	7.
2	2014	ACURA	ILX HYBRID	COMPACT	1.5	4	AV7	Z	6.0	5.
3	2014	ACURA	MDX 4WD	SUV - SMALL	3.5	6	AS6	Z	12.7	9.
4	2014	ACURA	RDX AWD	SUV - SMALL	3.5	6	AS6	Z	12.1	8.

5. Eksplorasi deskriptif data yang diunduh.

```
In [4]:
# merangkum data
df.describe()
```

Out[4]:

	MODELYEAR	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_CITY	FUELCONSUMPTION_HWY	FUELCONSUMPTION_COMB	FUELCONSUMPTION_COMB_
count	1067.0	1067.000000	1067.000000	1067.000000	1067.000000	1067.000000	1067.00
mean	2014.0	3.346298	5.794752	13.296532	9.474602	11.580881	26.44
std	0.0	1.415895	1.797447	4.101253	2.794510	3.485595	7.46
min	2014.0	1.000000	3.000000	4.600000	4.900000	4.700000	11.00
25%	2014.0	2.000000	4.000000	10.250000	7.500000	9.000000	21.00
50%	2014.0	3.400000	6.000000	12.600000	8.800000	10.900000	26.00
75%	2014.0	4.300000	8.000000	15.550000	10.850000	13.350000	31.00
max	2014.0	8.400000	12.000000	30.200000	20.500000	25.800000	60.00

6. Beberapa fitur dapat dieksplorasi dengan cara berikut.

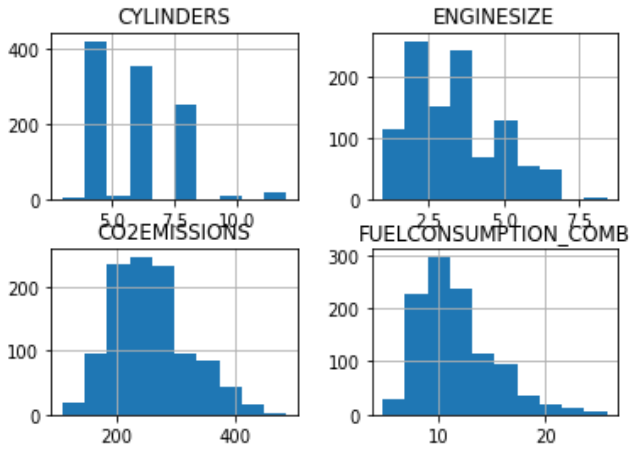
```
In [5]:
cdf = df[['ENGINESIZE','CYLINDERS','FUELCONSUMPTION_COMB','CO2EMISSIONS']]
cdf.head(10)
```

Out[5]:

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	212

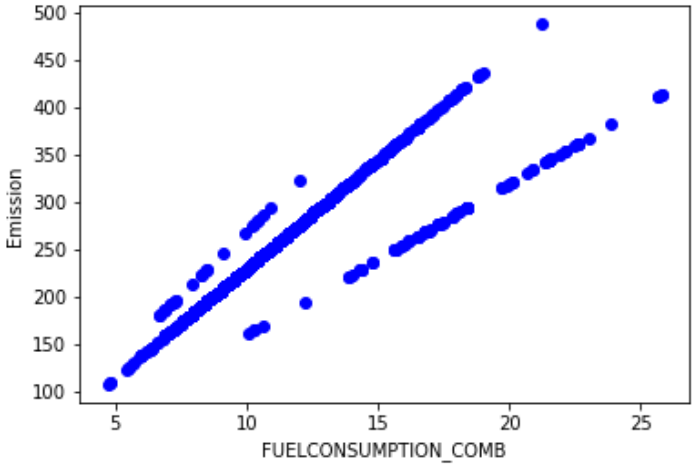
7. Fitur-fitur tersebut dapat diplot sebagai berikut:

```
In [6]:
viz = cdf[['CYLINDERS','ENGINESIZE','CO2EMISSIONS','FUELCONSUMPTION_COMB']]
viz.hist()
plt.show()
```

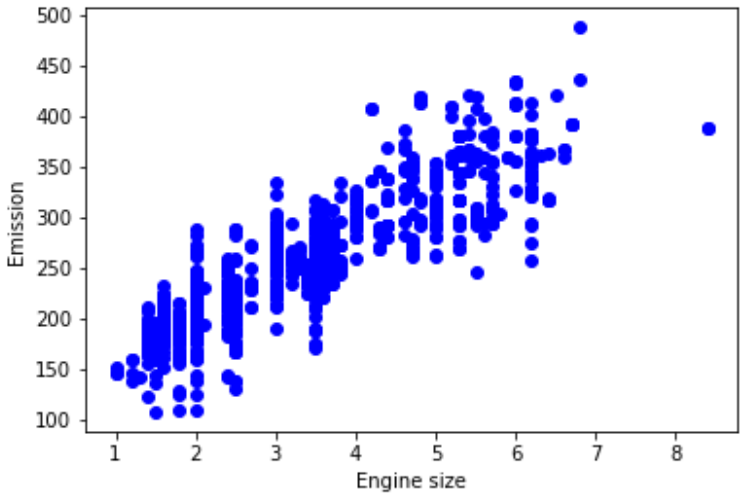


8. Plot fitur-fitur vs emisi dapat dibuat dan dapat dilihat linearitas hubungannya.

```
In [7]:
plt.scatter(cdf.FUELCONSUMPTION_COMB, cdf.CO2EMISSIONS,  color='blue')
plt.xlabel("FUELCONSUMPTION_COMB")
plt.ylabel("Emission")
plt.show()
```



In [8]:
```
plt.scatter(cdf.ENGINESIZE, cdf.CO2EMISSIONS,  color='blue')
plt.xlabel("Engine size")
plt.ylabel("Emission")
plt.show()
```



Challenge

1. Plot CYLINDER vs the Emission, untuk melihat hubungan linearnya:

Pembuatan dataset pelatihan dan pengujian

Pemisahan data latih/uji melibatkan pemisahan dataset menjadi dataset pelatihan dan pengujian, yang saling eksklusif. Setelah itu, dataset pelatihan dapat digunakan untuk membuat model dan dataset pengujian untuk pengujian. Hal ini akan memberikan evaluasi yang lebih akurat pada akurasi out-of-sample karena dataset pengujian bukan merupakan bagian dari dataset yang telah digunakan untuk melatih data. Ini lebih realistis untuk masalah dunia nyata.

Ini berarti bahwa hasil dari setiap titik data dalam kumpulan data ini diketahui, sehingga sangat bagus untuk data pengujian. Dataset pengujian belum digunakan untuk melatih model, sehingga model tidak memiliki pengetahuan tentang hasil dari data ini, sehingga dapat disebut pengujian di luar sampel

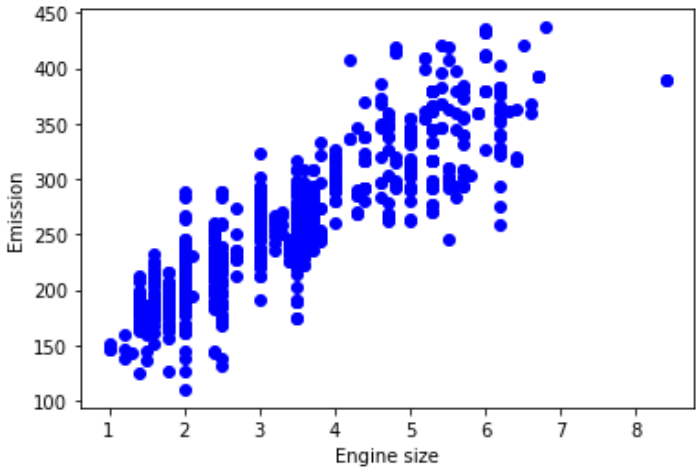
In [10]:
```
msk = np.random.rand(len(df)) < 0.8
train = cdf[msk]
test = cdf[~msk]
```

Model Regresi Sederhana

Regresi Linier cocok dengan model linier dengan koefisien B = (B1, ..., Bn) untuk meminimalkan 'jumlah kuadrat sisa (residu)' antara x bebas dalam kumpulan data, dan y tak bebas dengan pendekatan linier.

9. Distribusi Data Pelatihan.

```
In [11]:
plt.scatter(train.ENGINESIZE, train.CO2EMISSIONS,  color='blue')
plt.xlabel("Engine size")
plt.ylabel("Emission")
plt.show()
```



Pemodelan

10. Menggunakan sklearn package untuk memodelkan data.

```
In [12]:
from sklearn import linear_model
regr = linear_model.LinearRegression()
train_x = np.asanyarray(train[['ENGINESIZE']])
train_y = np.asanyarray(train[['CO2EMISSIONS']])
regr.fit (train_x, train_y)
# The coefficients
print ('Coefficients: ', regr.coef_)
print ('Intercept: ',regr.intercept_)
```

```
Coefficients:  [[39.01678181]]
Intercept:  [126.04158835]
```

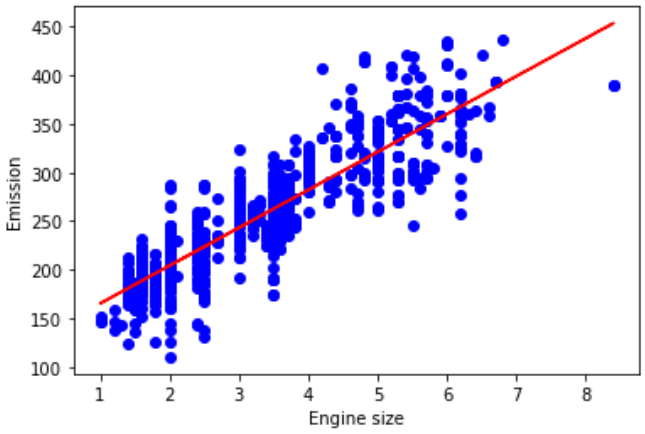
Seperti disebutkan sebelumnya, **koefisien** dan **intercept** dalam regresi linier sederhana, adalah parameter dari garis yang cocok dengan data. Mengingat bahwa ini adalah regresi linier sederhana, dengan hanya 2 parameter, dan mengetahui bahwa parameternya adalah intercept dan kemiringan atau gradien garis (koefisien), sklearn dapat memperkirakannya langsung dari data. Perhatikan bahwa semua data harus tersedia untuk menghitung parameter.

Plot output model

11. Plotting garis yang mencocoki terhadap data:

```
In [13]:
1 plt.scatter(train.ENGINESIZE, train.CO2EMISSIONS,  color='blue')
2 plt.plot(train_x, regr.coef_[0][0]*train_x + regr.intercept_[0], '-r')
3 plt.xlabel("Engine size")
4 plt.ylabel("Emission")
```

```
Out[13]: Text(0, 0.5, 'Emission')
```



Evaluasi

Nilai aktual dan nilai prediksi dapat dibandingkan untuk menghitung akurasi dari model regresi. Metrik evaluasi sangat penting untuk pengembangan model karena memberikan pengetahuan untuk perbaikan model.

Ada berbagai metrik untuk evaluasi model, misalnya MSE sebagai error untuk mengetahui akurasi dari model yang dibangun yang dihitung dari MSE model terhadap data pengujian: - Mean Absolute Error (MAE): Rerata dari nilai absolut dari error. MAE adalah metrik paling mudah dipahami karena hanya rata-rata dari error. - Mean Squared Error (MSE): adalah rerata dari error dikuadratkan. MSE lebih populer dibanding MAE karena fokus pada error yang besar karena dikuadratkan sehingga berdampak lebih besar terhadap error yang lebih besar dibandingkan error yang lebih kecil. - Root Mean Squared Error (RMSE). - R-squared bukan error namun metrik yang populer yang merepresentasikan sejauh mana data cocok dengan garis regresi yang didapatkan. Semakin besar R-squared akan semaki baik pencocokan garis terhadap data. Nilai terbaik adalah 1.0 dan dapat bernilai negatif.

```
In [14]:
from sklearn.metrics import r2_score

test_x = np.asanyarray(test[['ENGINESIZE']])
test_y = np.asanyarray(test[['CO2EMISSIONS']])
test_y_ = regr.predict(test_x)

print("Mean absolute error: %.2f" % np.mean(np.absolute(test_y_ - test_y)))
print("Residual sum of squares (MSE): %.2f" % np.mean((test_y_ - test_y) ** 2))
print("R2-score: %.2f" % r2_score(test_y_ , test_y) )
```

```
Mean absolute error: 24.33
Residual sum of squares (MSE): 1049.80
R2-score: 0.66
```

Model Regresi Variabel Jamak

1. Buatlah seperti langkah no. 9 untuk variabel – variabel: Engine size, Fuel Consumptions, dan Cylinders!
2. Lakukan langkah pemodelan seperti no.10 dengan menggunakan 3 variabel tersebut sebagai data X!
3. Lakukan sampai tahap evaluasinya!

PENGUMPULAN

1. File yang dikumpulkan terdiri dari:
    a. File project (.ipynb)
    b. File PDF berisi screenshot output dan jawaban (jika ada pertanyaan)
2. File di-compress (.zip) dan diberi nama **KODEMATAKULIAH_KELAS_NIM_NAMA_WEEK KE-XX.zip** (contoh: IS5411_A_13110310017_Monika Evelin Johan_Week-13.zip).

REFERENSI

Hall, M. A., Frank, E., Witten, I. H., Pal, C. (2016). Data Mining: Practical Machine Learning Tools and Techniques. Netherlands: Elsevier Science.