الجمهورية اليمنية

جامعة إب

كلية العلوم

قسم علوم الحاسوب وتقنية المعلومات

تكليف مقرر

تنقيب بيانات - عملي

Data Mining

للمحاضرتين الخامسة + السادسة

عمل الطالب :

أسامة سعيد محمد حمود سعيد - مجموعة A

إشراف :

أ. مالك المصنف

2024 - 2025

الخطوة 1

تحميل ال Dataset

```
from sklearn.datasets import load_iris
```

الخطوة 2

فصل اعمدة الميزات الى متغير وطباعته

```
datasets = load_iris()

X = datasets.data

X
```

```
array([[5.1, 3.5, 1.4, 0.2],
       [4.9, 3. , 1.4, 0.2],
       [4.7, 3.2, 1.3, 0.2],
       [4.6, 3.1, 1.5, 0.2],
       [5. , 3.6, 1.4, 0.2],
       [5.4, 3.9, 1.7, 0.4],
       [4.6, 3.4, 1.4, 0.3],
       [5. , 3.4, 1.5, 0.2],
       [4.4, 2.9, 1.4, 0.2],
       [4.9, 3.1, 1.5, 0.1],
       [5.4, 3.7, 1.5, 0.2],
       [4.8, 3.4, 1.6, 0.2],
       [4.8, 3. , 1.4, 0.1],
       [4.3, 3. , 1.1, 0.1],
       [5.8, 4. , 1.2, 0.2],
       [5.7, 4.4, 1.5, 0.4],
       [5.4, 3.9, 1.3, 0.4],
       [5.1, 3.5, 1.4, 0.3],
       [5.7, 3.8, 1.7, 0.3],
       [5.1, 3.8, 1.5, 0.3],
       [5.4, 3.4, 1.7, 0.2],
       [5.1, 3.7, 1.5, 0.4],
       [4.6, 3.6, 1. , 0.2],
       [5.1, 3.3, 1.7, 0.5],
       [4.8, 3.4, 1.9, 0.2],
       [5. , 3. , 1.6, 0.2],
       [5. , 3.4, 1.6, 0.4],
       [5.2, 3.5, 1.5, 0.2],
       [5.2, 3.4, 1.4, 0.2],
       [4.7, 3.2, 1.6, 0.2],
       [4.8, 3.1, 1.6, 0.2],
```

```
       [5.4, 3.4, 1.5, 0.4],
       [5.2, 4.1, 1.5, 0.1],
       [5.5, 4.2, 1.4, 0.2],
       [4.9, 3.1, 1.5, 0.2],
       [5. , 3.2, 1.2, 0.2],
       [5.5, 3.5, 1.3, 0.2],
       [4.9, 3.6, 1.4, 0.1],
       [4.4, 3. , 1.3, 0.2],
       [5.1, 3.4, 1.5, 0.2],
       [5. , 3.5, 1.3, 0.3],
       [4.5, 2.3, 1.3, 0.3],
       [4.4, 3.2, 1.3, 0.2],
       [5. , 3.5, 1.6, 0.6],
       [5.1, 3.8, 1.9, 0.4],
       [4.8, 3. , 1.4, 0.3],
       [5.1, 3.8, 1.6, 0.2],
       [4.6, 3.2, 1.4, 0.2],
       [5.3, 3.7, 1.5, 0.2],
       [5. , 3.3, 1.4, 0.2],
       [7. , 3.2, 4.7, 1.4],
       [6.4, 3.2, 4.5, 1.5],
       [6.9, 3.1, 4.9, 1.5],
       [5.5, 2.3, 4. , 1.3],
       [6.5, 2.8, 4.6, 1.5],
       [5.7, 2.8, 4.5, 1.3],
       [6.3, 3.3, 4.7, 1.6],
       [4.9, 2.4, 3.3, 1. ],
       [6.6, 2.9, 4.6, 1.3],
       [5.2, 2.7, 3.9, 1.4],
       [5. , 2. , 3.5, 1. ],
       [5.9, 3. , 4.2, 1.5],
       [6. , 2.2, 4. , 1. ],
       [6.1, 2.9, 4.7, 1.4],
       [5.6, 2.9, 3.6, 1.3],
       [6.7, 3.1, 4.4, 1.4],
       [5.6, 3. , 4.5, 1.5],
       [5.8, 2.7, 4.1, 1. ],
       [6.2, 2.2, 4.5, 1.5],
       [5.6, 2.5, 3.9, 1.1],
       [5.9, 3.2, 4.8, 1.8],
       [6.1, 2.8, 4. , 1.3],
       [6.3, 2.5, 4.9, 1.5],
       [6.1, 2.8, 4.7, 1.2],
       [6.4, 2.9, 4.3, 1.3],
       [6.6, 3. , 4.4, 1.4],
       [6.8, 2.8, 4.8, 1.4],
       [6.7, 3. , 5. , 1.7],
       [6. , 2.9, 4.5, 1.5],
       [5.7, 2.6, 3.5, 1. ],
```

```
       [5.5, 2.4, 3.8, 1.1],
       [5.5, 2.4, 3.7, 1. ],
       [5.8, 2.7, 3.9, 1.2],
       [6. , 2.7, 5.1, 1.6],
       [5.4, 3. , 4.5, 1.5],
       [6. , 3.4, 4.5, 1.6],
       [6.7, 3.1, 4.7, 1.5],
       [6.3, 2.3, 4.4, 1.3],
       [5.6, 3. , 4.1, 1.3],
       [5.5, 2.5, 4. , 1.3],
       [5.5, 2.6, 4.4, 1.2],
       [6.1, 3. , 4.6, 1.4],
       [5.8, 2.6, 4. , 1.2],
       [5. , 2.3, 3.3, 1. ],
       [5.6, 2.7, 4.2, 1.3],
       [5.7, 3. , 4.2, 1.2],
       [5.7, 2.9, 4.2, 1.3],
       [6.2, 2.9, 4.3, 1.3],
       [5.1, 2.5, 3. , 1.1],
       [5.7, 2.8, 4.1, 1.3],
       [6.3, 3.3, 6. , 2.5],
       [5.8, 2.7, 5.1, 1.9],
       [7.1, 3. , 5.9, 2.1],
       [6.3, 2.9, 5.6, 1.8],
       [6.5, 3. , 5.8, 2.2],
       [7.6, 3. , 6.6, 2.1],
       [4.9, 2.5, 4.5, 1.7],
       [7.3, 2.9, 6.3, 1.8],
       [6.7, 2.5, 5.8, 1.8],
       [7.2, 3.6, 6.1, 2.5],
       [6.5, 3.2, 5.1, 2. ],
       [6.4, 2.7, 5.3, 1.9],
       [6.8, 3. , 5.5, 2.1],
       [5.7, 2.5, 5. , 2. ],
       [5.8, 2.8, 5.1, 2.4],
       [6.4, 3.2, 5.3, 2.3],
       [6.5, 3. , 5.5, 1.8],
       [7.7, 3.8, 6.7, 2.2],
       [7.7, 2.6, 6.9, 2.3],
       [6. , 2.2, 5. , 1.5],
       [6.9, 3.2, 5.7, 2.3],
       [5.6, 2.8, 4.9, 2. ],
       [7.7, 2.8, 6.7, 2. ],
       [6.3, 2.7, 4.9, 1.8],
       [6.7, 3.3, 5.7, 2.1],
       [7.2, 3.2, 6. , 1.8],
       [6.2, 2.8, 4.8, 1.8],
       [6.1, 3. , 4.9, 1.8],
       [6.4, 2.8, 5.6, 2.1],
```

```
       [7.2, 3. , 5.8, 1.6],
       [7.4, 2.8, 6.1, 1.9],
       [7.9, 3.8, 6.4, 2. ],
       [6.4, 2.8, 5.6, 2.2],
       [6.3, 2.8, 5.1, 1.5],
       [6.1, 2.6, 5.6, 1.4],
       [7.7, 3. , 6.1, 2.3],
       [6.3, 3.4, 5.6, 2.4],
       [6.4, 3.1, 5.5, 1.8],
       [6. , 3. , 4.8, 1.8],
       [6.9, 3.1, 5.4, 2.1],
       [6.7, 3.1, 5.6, 2.4],
       [6.9, 3.1, 5.1, 2.3],
       [5.8, 2.7, 5.1, 1.9],
       [6.8, 3.2, 5.9, 2.3],
       [6.7, 3.3, 5.7, 2.5],
       [6.7, 3. , 5.2, 2.3],
       [6.3, 2.5, 5. , 1.9],
       [6.5, 3. , 5.2, 2. ],
       [6.2, 3.4, 5.4, 2.3],
       [5.9, 3. , 5.1, 1.8]])
```

طباعة اسماء الاعمدة

```
datasets.feature_names
```

```
['sepal length (cm)',
 'sepal width (cm)',
 'petal length (cm)',
 'petal width (cm)']
```

استخراج العمود الهدف الى متغير اخر وطباعته

```
Y = datasets.target
```

```
Y
```

```
array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
       2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
       2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2])
```

طباعه اسماء الكلاسات التصنيفية الخاصة بالداتا

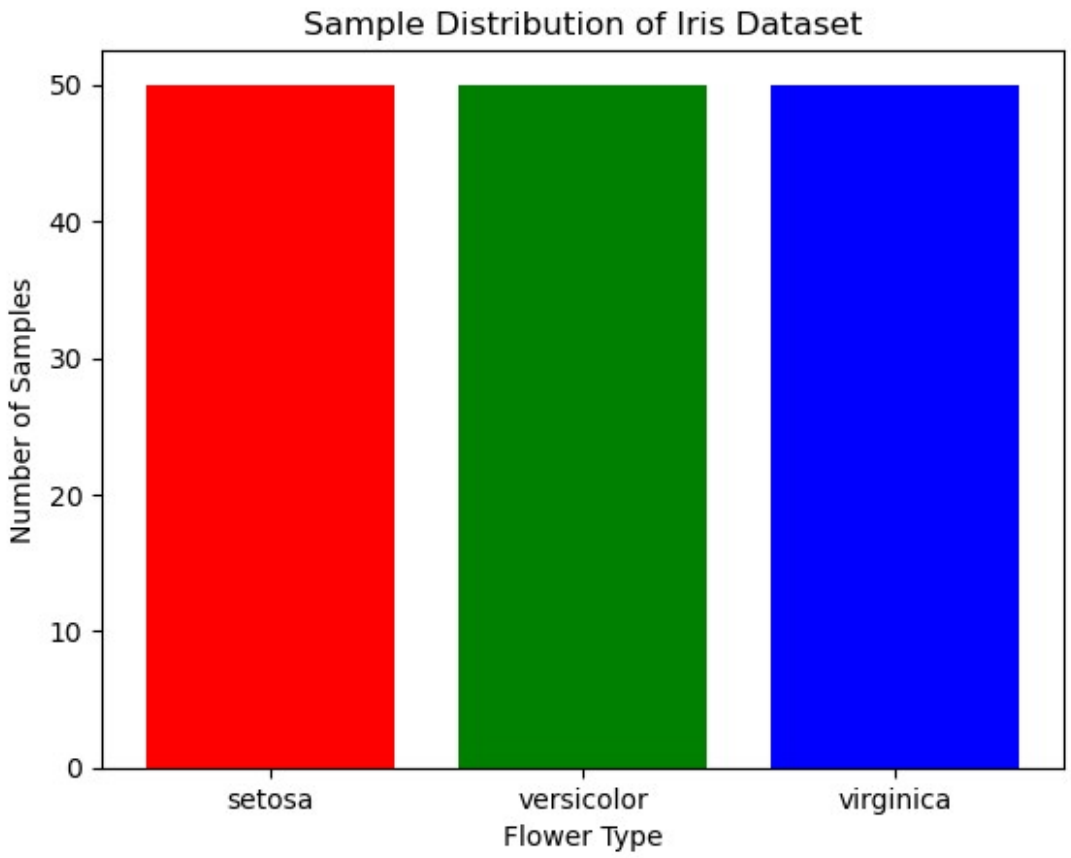
```
datasets.target_names
```

```
array(['setosa', 'versicolor', 'virginica'], dtype='<U10')
```

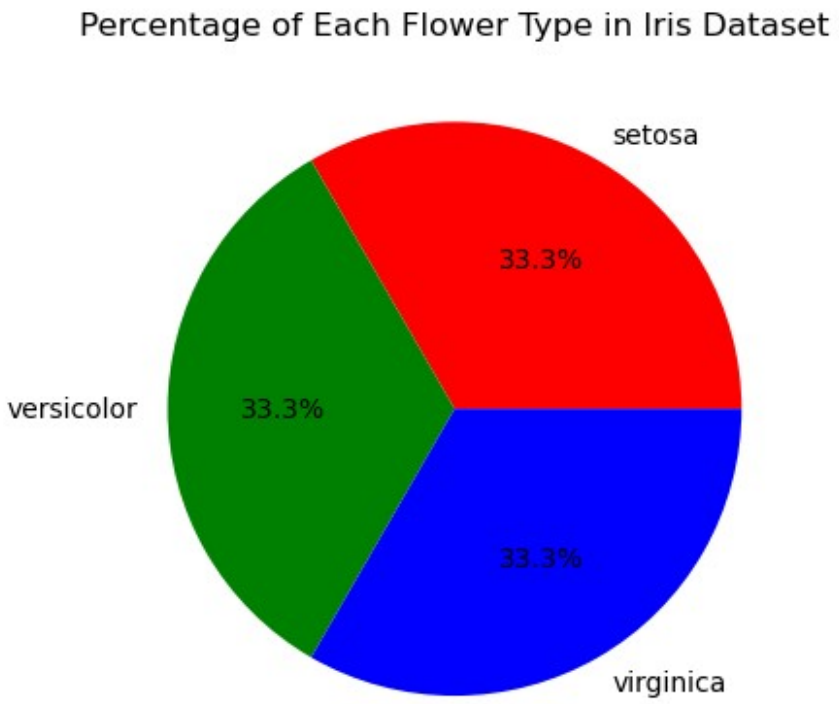
الخطوة 3

تمثيل الداتا على شكل bar **و** pie **في الموجودة** Matplotlib

```
import matplotlib.pyplot as plt
import numpy as np

target_names = datasets.target_names
unique, counts = np.unique(Y, return_counts=True)
plt.bar(target_names, counts, color=['red', 'green', 'blue'])
plt.xlabel("Flower Type")
plt.ylabel("Number of Samples")
plt.title("Sample Distribution of Iris Dataset")
plt.show()
```



```
plt.pie(counts, labels=target_names, autopct='%1.1f%%', colors=['red',
'green', 'blue'])
plt.title("Percentage of Each Flower Type in Iris Dataset")
plt.show()
```



4 الخطوة

قراءة الداتا الخاصة بسيارات TOYOTA

```
import pandas as pd

dataset_car = pd.read_csv('dataintegration.csv')

dataset_car
```

	Unnamed: 0	Car_Price	Vehicle_Age	KM_Travelled	Fuel_Type	HP
0	0	13500	23	46986	0	90
1	1	13750	23	72937	0	90
2	2	13950	24	41711	0	90
3	3	14950	26	48000	0	90

```
4              4      13750           30         38500           0   90
...          ...        ...          ...           ...         ...  ...
1431        1431       7500           47         20544           1   86
1432        1432      10845           47         11000           1   86
1433        1433       8500           47         17016           1   86
1434        1434       7250           47         11000           1   86
1435        1435       6950           47             1           1  110
```

```
        Paint_Type Transmission_Type  Engine_Size  Doors  Weight
0         Metallic            Manual         2000      3    1165
1         Metallic            Manual         2000      3    1165
2         Metallic            Manual         2000      3    1165
3     Non-Metallic            Manual         2000      3    1165
4     Non-Metallic            Manual         2000      3    1170
...            ...               ...          ...    ...     ...
1431      Metallic            Manual         1300      3    1025
1432  Non-Metallic            Manual         1300      3    1015
1433  Non-Metallic            Manual         1300      3    1015
1434      Metallic            Manual         1300      3    1015
1435  Non-Metallic            Manual         1600      5    1114

[1436 rows x 11 columns]
```

```
dataset_car.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1436 entries, 0 to 1435
Data columns (total 11 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Unnamed: 0         1436 non-null   int64
 1   Car_Price          1436 non-null   int64
 2   Vehicle_Age        1436 non-null   int64
 3   KM_Travelled       1436 non-null   int64
 4   Fuel_Type          1436 non-null   int64
 5   HP                 1436 non-null   int64
 6   Paint_Type         1436 non-null   object
 7   Transmission_Type  1436 non-null   object
 8   Engine_Size        1436 non-null   int64
 9   Doors              1436 non-null   int64
 10  Weight             1436 non-null   int64
dtypes: int64(9), object(2)
memory usage: 123.5+ KB
```

```
 dataset_car.isna().sum()
```

```
Unnamed: 0           0
Car_Price            0
Vehicle_Age          0
KM_Travelled         0
Fuel_Type            0
HP                   0
Paint_Type           0
Transmission_Type    0
Engine_Size          0
Doors                0
Weight               0
dtype: int64
```

تحويل الاعمدة التي من نوع object **إلى** int **بأحد أساليب الترميز** Encoders **التي تمت دراستها**

```
dataset_car.Paint_Type.unique()
```

```
array(['Metallic', 'Non-Metallic'], dtype=object)
```

```
dataset_car.Transmission_Type.unique()
```

```
array(['Manual', 'Auto'], dtype=object)
```

```
from sklearn.preprocessing import LabelEncoder

Encoder = LabelEncoder()

dataset_car.Paint_Type = Encoder.fit_transform(dataset_car.Paint_Type)


dataset_car.Paint_Type.unique()
```

```
array([0, 1], dtype=int64)
```

```
dataset_car.Transmission_Type =
Encoder.fit_transform(dataset_car.Transmission_Type)


dataset_car.Transmission_Type.unique()
```

```
array([1, 0], dtype=int64)
```

```
dataset_car.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1436 entries, 0 to 1435
Data columns (total 11 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
```

```
 0   Unnamed: 0          1436 non-null   int64
 1   Car_Price           1436 non-null   int64
 2   Vehicle_Age         1436 non-null   int64
 3   KM_Travelled        1436 non-null   int64
 4   Fuel_Type           1436 non-null   int64
 5   HP                  1436 non-null   int64
 6   Paint_Type          1436 non-null   int64
 7   Transmission_Type   1436 non-null   int64
 8   Engine_Size         1436 non-null   int64
 9   Doors               1436 non-null   int64
 10  Weight              1436 non-null   int64
dtypes: int64(11)
memory usage: 123.5 KB
```

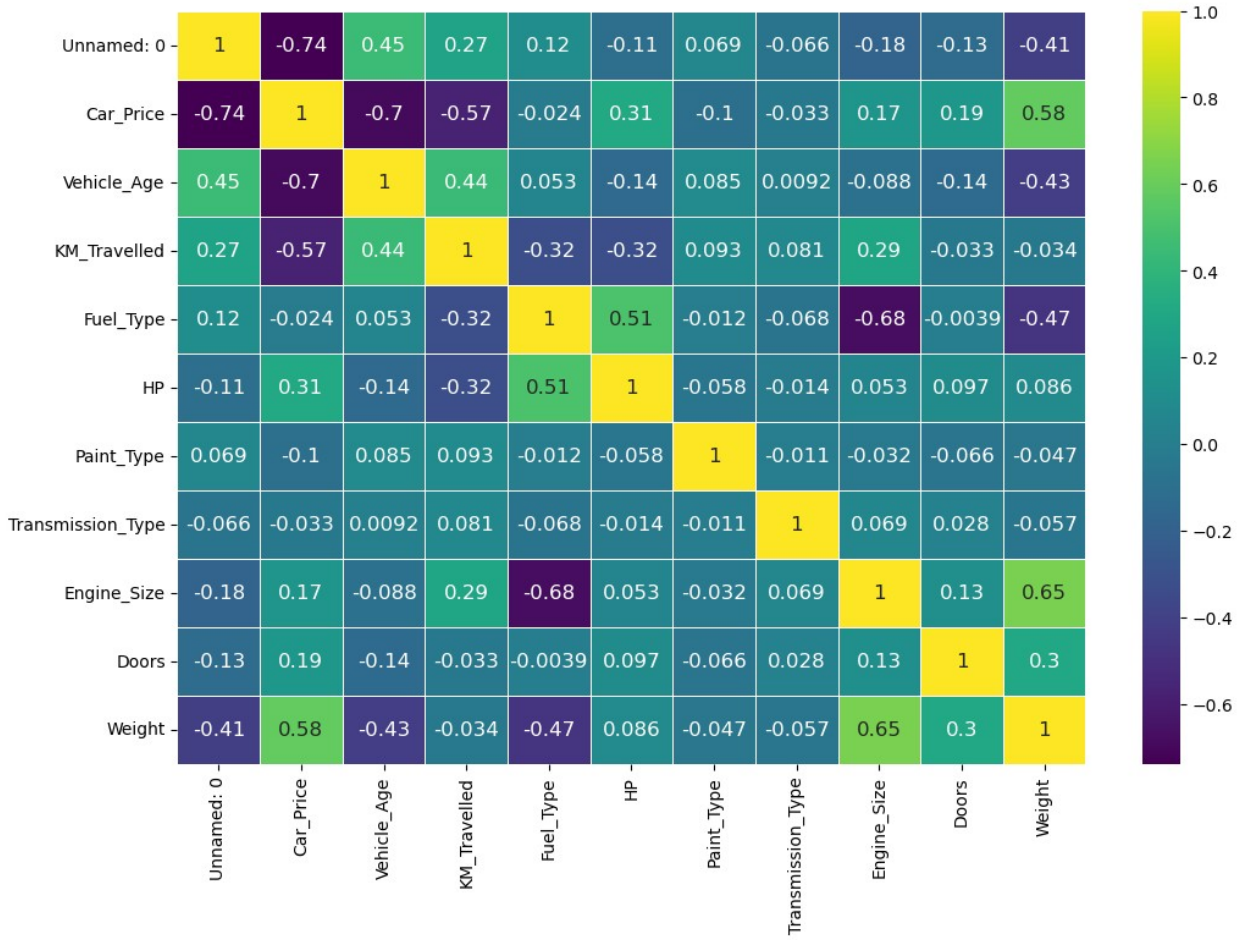
```
dataset_car.describe()
```

```
        Unnamed: 0     Car_Price  Vehicle_Age   KM_Travelled
Fuel_Type  \
count  1436.000000   1436.000000  1436.000000    1436.000000
1436.000000
mean    717.500000  10730.824513    47.476323   68045.075209
0.910167
std     414.681806   3626.964585    13.306889   37597.343766
0.320509
min       0.000000   4350.000000     1.000000       1.000000
0.000000
25%     358.750000   8450.000000    44.000000   42702.500000
1.000000
50%     717.500000   9900.000000    47.000000   63061.500000
1.000000
75%    1076.250000  11950.000000    55.000000   86916.000000
1.000000
max    1435.000000  32500.000000    68.000000  243000.000000
2.000000

                HP   Paint_Type  Transmission_Type  Engine_Size
Doors  \
count  1436.000000  1436.000000        1436.000000  1436.000000
1436.000000
mean    101.476323     0.291086           0.944290  1566.827994
4.033426
std      14.737380     0.454421           0.229441   187.182436
0.952677
min      69.000000     0.000000           0.000000  1300.000000
2.000000
25%      90.000000     0.000000           1.000000  1400.000000
3.000000
50%     110.000000     0.000000           1.000000  1600.000000
4.000000
75%     110.000000     1.000000           1.000000  1600.000000
```

```
5.000000
max     192.000000     1.000000           1.000000  2000.000000
5.000000

           Weight
count  1436.00000
mean   1072.45961
std      52.64112
min    1000.00000
25%    1040.00000
50%    1070.00000
75%    1085.00000
max    1615.00000
```

رسم heatmap **للعلاقات بين الاعمدة الخاصة بها بحسب قانون** person

```
import seaborn as sns

correlation_matrix = dataset_car.corr()

plt.figure(figsize = (12,8))
sns.heatmap(
    correlation_matrix,
    annot = True,
    cmap = 'viridis',
    annot_kws = {"size" : 12},
    linewidth = 0.5,
    linecolor = 'white'
)
plt.show()
```


Unnamed: 0 1 -0.74 0.45 0.27 0.12 -0.11 0.069 -0.066 -0.18 -0.13 -0.41
Car_Price -0.74 1 -0.7 -0.57 -0.024 0.31 -0.1 -0.033 0.17 0.19 0.58
Vehicle_Age 0.45 -0.7 1 0.44 0.053 -0.14 0.085 0.0092 -0.088 -0.14 -0.43
KM_Travelled 0.27 -0.57 0.44 1 -0.32 -0.32 0.093 0.081 0.29 -0.033 -0.034
Fuel_Type 0.12 -0.024 0.053 -0.32 1 0.51 -0.012 -0.068 -0.68 -0.0039 -0.47
HP -0.11 0.31 -0.14 -0.32 0.51 1 -0.058 -0.014 0.053 0.097 0.086
Paint_Type 0.069 -0.1 0.085 0.093 -0.012 -0.058 1 -0.011 -0.032 -0.066 -0.047
Transmission_Type -0.066 -0.033 0.0092 0.081 -0.068 -0.014 -0.011 1 0.069 0.028 -0.057
Engine_Size -0.18 0.17 -0.088 0.29 -0.68 0.053 -0.032 0.069 1 0.13 0.65
Doors -0.13 0.19 -0.14 -0.033 -0.0039 0.097 -0.066 0.028 0.13 1 0.3
Weight -0.41 0.58 -0.43 -0.034 -0.47 0.086 -0.047 -0.057 0.65 0.3 1
Unnamed: 0
Car_Price
Vehicle_Age
KM_Travelled
Fuel_Type
HP
Paint_Type
Transmission_Type
Engine_Size
Doors
Weight
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6