Jour 41 : Housing Boston

L'objectif de ce cas pratique est de *__prédire le prix médian (medv) de logements à Boston__*. Le Boston Housing Dataset est un jeu de données classique en machine learning et statistique. Il décrit le marché immobilier de Boston, en ciblant les facteurs clés des prix (criminalité, taxes, éducation, accès aux autoroutes, etc.). Basé sur des données de recensement, il permet d'analyser des tendances, de créer des modèles prédictifs et de comprendre les dynamiques des marchés urbainsoccupées par leur propriétaire en milliers de dollars (medv).

0. Chargement des librairies

In [46]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score
```

1. Chargement de la base de données

In [12]:
```
import pandas as pd

data = pd.read_csv("data/BostonHousing.csv")
data.head()
```

Out[12]:

	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	b	lst
0	0.00632	18.0	2.31	0	0.538	6.575	65.2	4.0900	1	296	15.3	396.90	4
1	0.02731	0.0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.90	9
2	0.02729	0.0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4
3	0.03237	0.0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	2
4	0.06905	0.0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	396.90	5

2. Identification et traitement des valeurs manquantes (Proportion)

In [14]:
```
data.isnull().mean() * 100
```

Out[14]:
```
crim       0.000000
zn         0.000000
indus      0.000000
chas       0.000000
nox        0.000000
rm         0.988142
age        0.000000
dis        0.000000
rad        0.000000
tax        0.000000
ptratio    0.000000
b          0.000000
lstat      0.000000
medv       0.000000
dtype: float64
```

Notre base de données ne contiennent aucune valeur manquante, à l'exception du nombre moyen de pièces par logement (rm) qui contient 0.98% de valeurs manquantes.

In [15]:
```
# Imputation par la moyenne
data["rm"] = data["rm"].fillna(data["rm"].mean())
```

In [16]:
```
# Vérification
data.isnull().mean() * 100
```

Out[16]:
```
crim       0.0
zn         0.0
indus      0.0
chas       0.0
nox        0.0
rm         0.0
age        0.0
dis        0.0
rad        0.0
tax        0.0
ptratio    0.0
b          0.0
lstat      0.0
medv       0.0
dtype: float64
```

3. Analyse descriptive

3.1. Statistiques descriptives

In [18]:
```
data.describe().transpose()
```

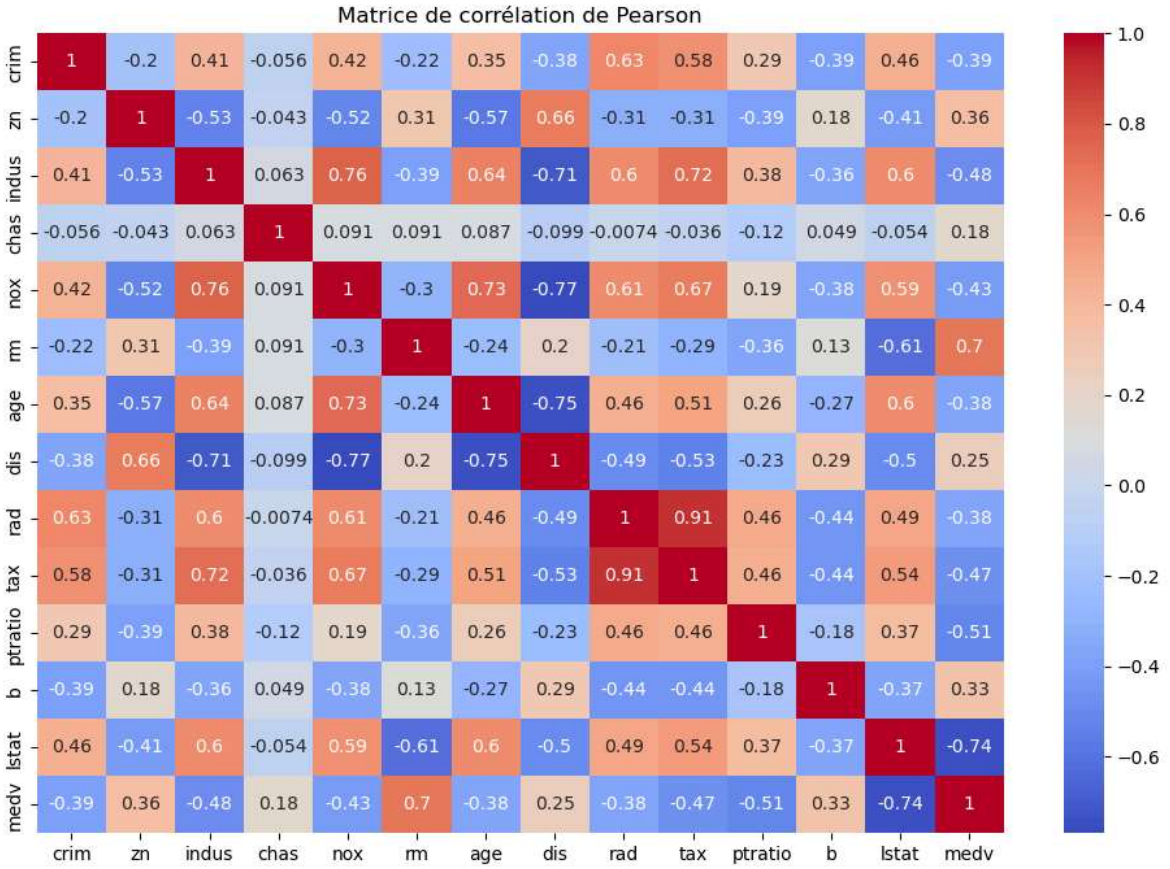
Out[18]:

	count	mean	std	min	25%	50%	75%
crim	506.0	3.613524	8.601545	0.00632	0.082045	0.25651	3.677083
zn	506.0	11.363636	23.322453	0.00000	0.000000	0.00000	12.500000
indus	506.0	11.136779	6.860353	0.46000	5.190000	9.69000	18.100000
chas	506.0	0.069170	0.253994	0.00000	0.000000	0.00000	0.000000
nox	506.0	0.554695	0.115878	0.38500	0.449000	0.53800	0.624000
rm	506.0	6.284341	0.702085	3.56100	5.885500	6.21000	6.618750
age	506.0	68.574901	28.148861	2.90000	45.025000	77.50000	94.075000
dis	506.0	3.795043	2.105710	1.12960	2.100175	3.20745	5.188425
rad	506.0	9.549407	8.707259	1.00000	4.000000	5.00000	24.000000
tax	506.0	408.237154	168.537116	187.00000	279.000000	330.00000	666.000000
ptratio	506.0	18.455534	2.164946	12.60000	17.400000	19.05000	20.200000
b	506.0	356.674032	91.294864	0.32000	375.377500	391.44000	396.225000
lstat	506.0	12.653063	7.141062	1.73000	6.950000	11.36000	16.955000
medv	506.0	22.532806	9.197104	5.00000	17.025000	21.20000	25.000000

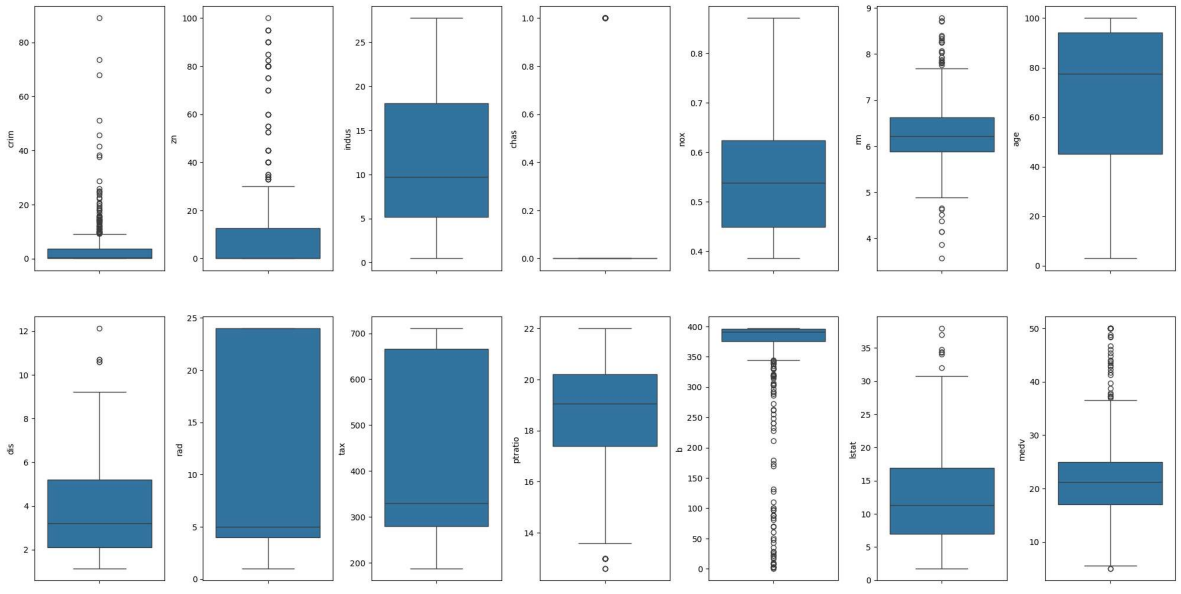
3.2. Analyse de la matrice de corrélation linéaire de Pearson

In [42]:

```
correlation_matrix = data.corr()

plt.figure(figsize=(12, 8))
sns.heatmap(
    correlation_matrix,
    annot=True,  # Afficher les valeurs
    cmap='coolwarm',  # Palette de couleurs
)

plt.title("Matrice de corrélation de Pearson")
plt.show()
```



In [49]:
```
# Création de boxplots
fig, ax = plt.subplots(ncols=7, nrows=2, figsize=(20, 10))
index = 0
ax = ax.flatten()

for col, value in data.items():
    sns.boxplot(y=col, data=data, ax=ax[index])
    index += 1
plt.tight_layout(pad=0.5, w_pad=0.7, h_pad=5.0)
```



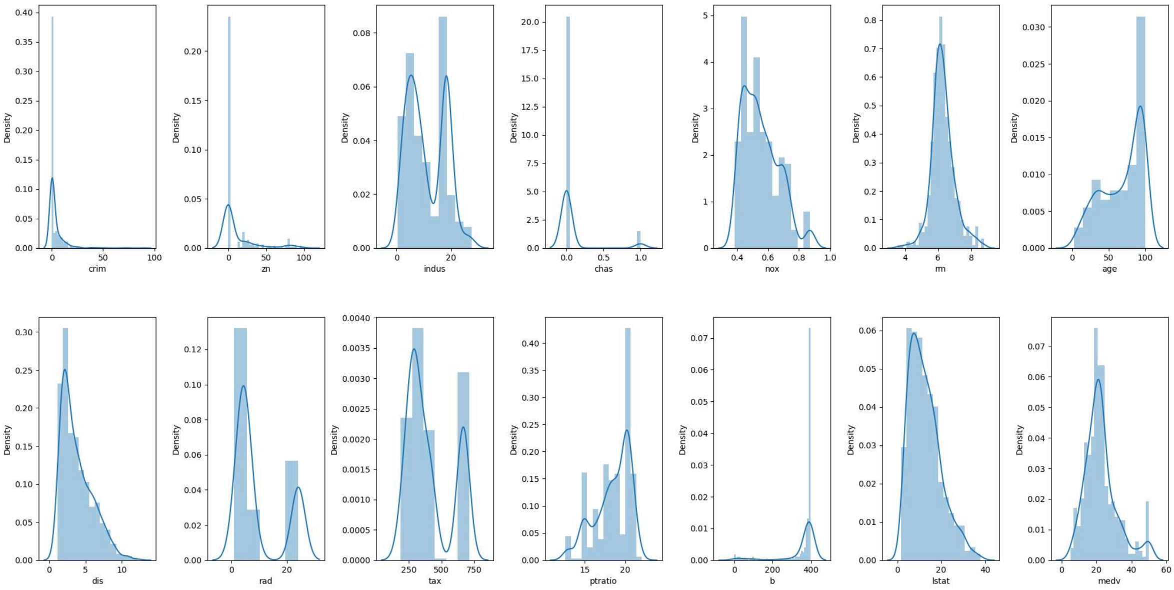
In [52]:
```
# Densités de distribution des données
fig, ax = plt.subplots(ncols=7, nrows=2, figsize=(20, 10))
index = 0
ax = ax.flatten()
```

```
for col, value in data.items():
    sns.distplot(value, ax=ax[index])
    index += 1
plt.tight_layout(pad=0.5, w_pad=0.7, h_pad=5.0)
plt.show()
```

```
C:\Users\HP\AppData\Local\Temp\ipykernel_21956\1050115955.py:7: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(value, ax=ax[index])
C:\Users\HP\AppData\Local\Temp\ipykernel_21956\1050115955.py:7: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(value, ax=ax[index])
C:\Users\HP\AppData\Local\Temp\ipykernel_21956\1050115955.py:7: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(value, ax=ax[index])
C:\Users\HP\AppData\Local\Temp\ipykernel_21956\1050115955.py:7: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(value, ax=ax[index])
C:\Users\HP\AppData\Local\Temp\ipykernel_21956\1050115955.py:7: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(value, ax=ax[index])
C:\Users\HP\AppData\Local\Temp\ipykernel_21956\1050115955.py:7: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
```

```
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(value, ax=ax[index])
C:\Users\HP\AppData\Local\Temp\ipykernel_21956\1050115955.py:7: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(value, ax=ax[index])
C:\Users\HP\AppData\Local\Temp\ipykernel_21956\1050115955.py:7: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(value, ax=ax[index])
C:\Users\HP\AppData\Local\Temp\ipykernel_21956\1050115955.py:7: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(value, ax=ax[index])
C:\Users\HP\AppData\Local\Temp\ipykernel_21956\1050115955.py:7: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(value, ax=ax[index])
C:\Users\HP\AppData\Local\Temp\ipykernel_21956\1050115955.py:7: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751
```

```
  sns.distplot(value, ax=ax[index])
C:\Users\HP\AppData\Local\Temp\ipykernel_21956\1050115955.py:7: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(value, ax=ax[index])
C:\Users\HP\AppData\Local\Temp\ipykernel_21956\1050115955.py:7: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(value, ax=ax[index])
C:\Users\HP\AppData\Local\Temp\ipykernel_21956\1050115955.py:7: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(value, ax=ax[index])
```



4. Modélisation

4.1. Préparation des données

In [30]:
```
X = data.drop('medv', axis=1)
y = data['medv']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_
```

In [32]:
```
# Modélisation
model = LinearRegression()
model.fit(X_train, y_train)
```

Out[32]:
```
▾ LinearRegression

LinearRegression()
```

In [ ]:
```
# Validation du modèle
```

In [33]:
```
# Prédiction
predictions = model.predict(X_test)
```

In [47]:
```
# Évaluation
r2 = r2_score(y_test, predictions)
print(f"R²: {r2}")
```

```
R²: 0.6672089705941917
```

In [ ]: