In [12]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

In [30]:
```
%matplotlib inline
import warnings
warnings.filterwarnings(action="ignore")
df=pd.read_csv("/home/student/Downloads/housing.csv")
df.head()
```

Out[30]:

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
0	0.00632	18.0	2.31	0.0	0.538	6.575	65.2	4.0900	1	296	15.3	396.90	4.98	24.0
1	0.02731	0.0	7.07	0.0	0.469	6.421	78.9	4.9671	2	242	17.8	396.90	9.14	21.6
2	0.02729	0.0	7.07	0.0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
3	0.03237	0.0	2.18	0.0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4
4	0.06905	0.0	2.18	0.0	0.458	7.147	54.2	6.0622	3	222	18.7	396.90	NaN	36.2

In [31]:
```
df.shape
```

Out[31]:
```
(506, 14)
```


In [32]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 506 entries, 0 to 505
Data columns (total 14 columns):
 #   Column   Non-Null Count  Dtype  
---  ------   --------------  -----  
 0   CRIM     486 non-null    float64
 1   ZN       486 non-null    float64
 2   INDUS    486 non-null    float64
 3   CHAS     486 non-null    float64
 4   NOX      506 non-null    float64
 5   RM       506 non-null    float64
 6   AGE      486 non-null    float64
 7   DIS      506 non-null    float64
 8   RAD      506 non-null    int64  
 9   TAX      506 non-null    int64  
 10  PTRATIO  506 non-null    float64
 11  B        506 non-null    float64
 12  LSTAT    486 non-null    float64
 13  MEDV     506 non-null    float64
dtypes: float64(12), int64(2)
memory usage: 55.5 KB
```

In [33]:
```
df.isnull().sum()
```

Out[33]:
```
CRIM       20
ZN         20
INDUS      20
CHAS       20
NOX         0
RM          0
AGE        20
DIS         0
RAD         0
TAX         0
PTRATIO     0
B           0
LSTAT      20
MEDV        0
dtype: int64
```

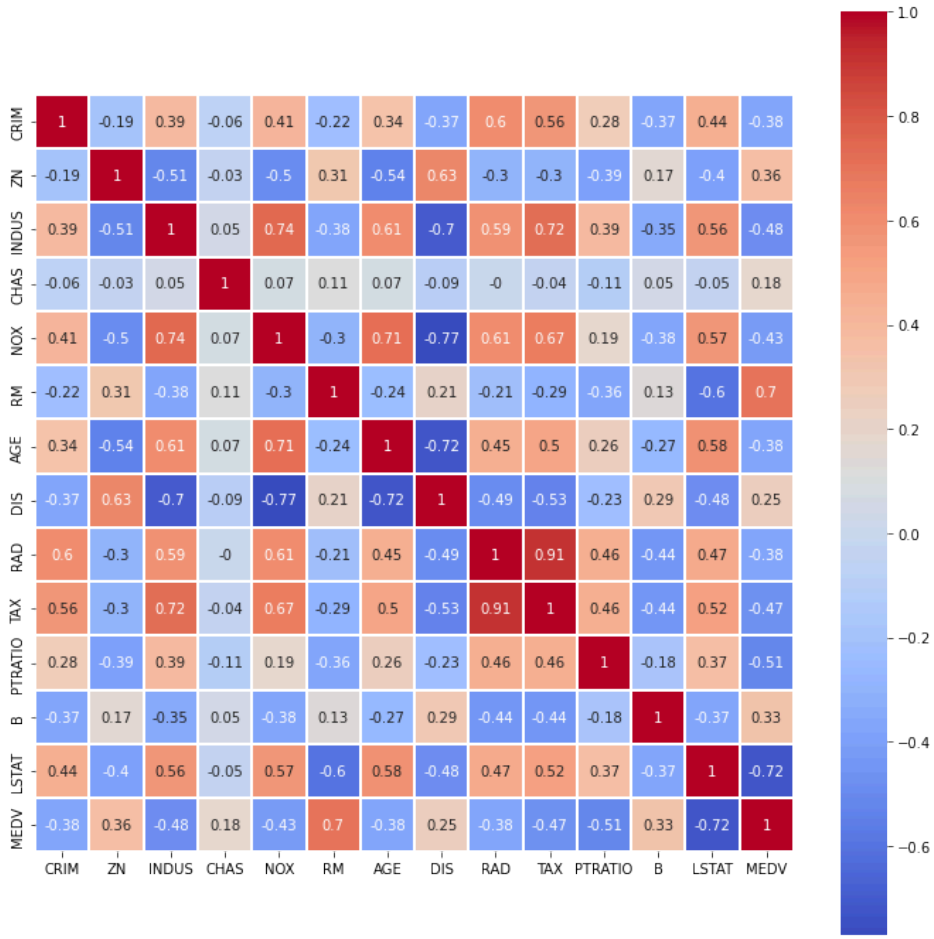
In [34]:
```
name=["CRIM","ZN","INDUS","CHAS","NOX","RM","AGE","DIS","RAD","TAX","PTRATIO","B","LSTAT","MEDV"]
for i in name:
    df[i].fillna(df[i].median(),inplace=True)
```

In [35]:
```
df.isnull().sum()
```

Out[35]:
```
CRIM       0
ZN         0
INDUS      0
CHAS       0
NOX        0
RM         0
AGE        0
DIS        0
RAD        0
TAX        0
PTRATIO    0
B          0
LSTAT      0
MEDV       0
dtype: int64
```


In [37]:
```
plt.figure(figsize=(12,12))
sns.heatmap(data=df.corr().round(2),annot=True,cmap='coolwarm',linewidths=0.2,square=True)
```

Out[37]:
```
<Axes: >
```



In [49]:
```
df1=df[['RM','LSTAT','PTRATIO','TAX','MEDV']]
df1
```

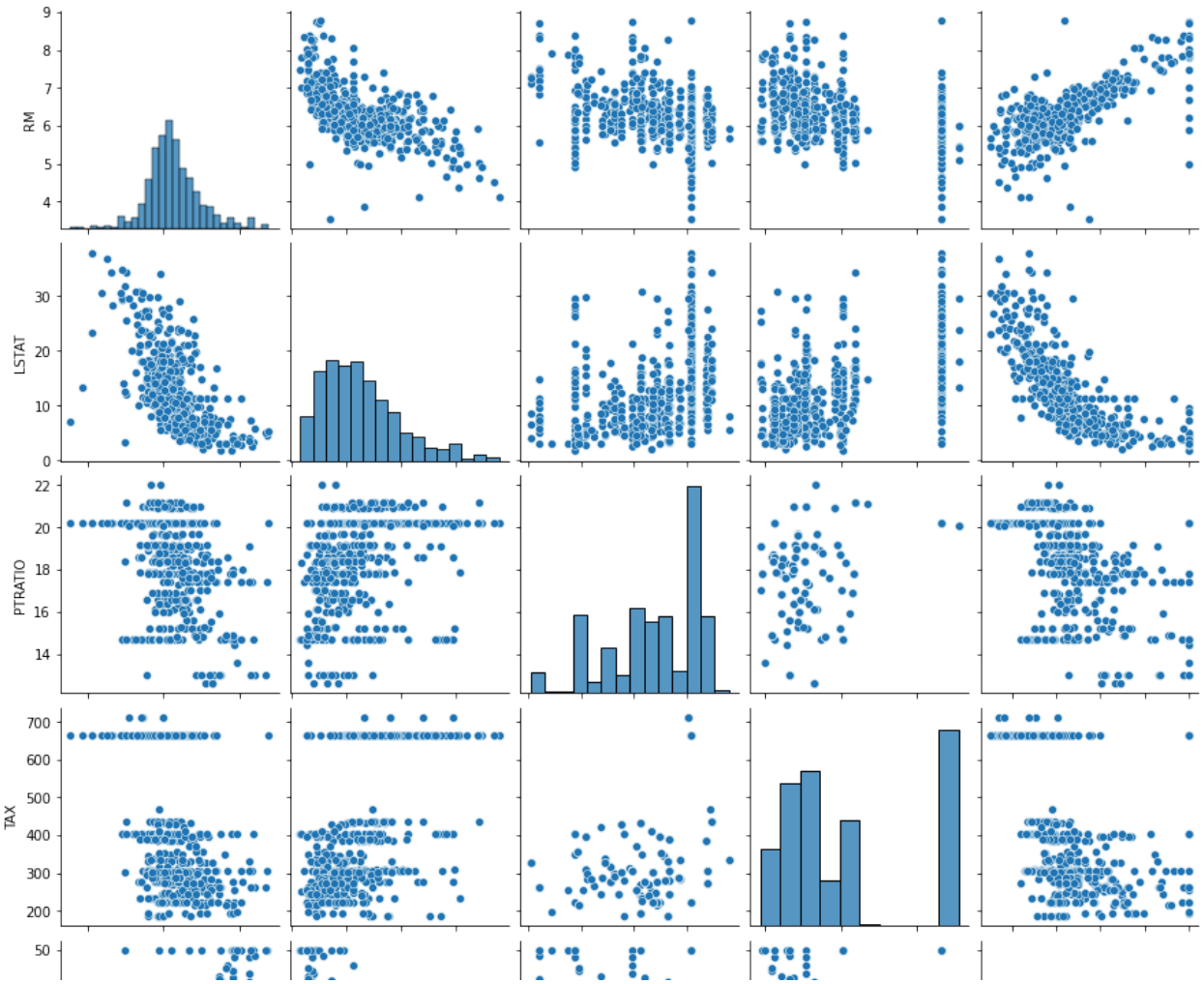
Out[49]:

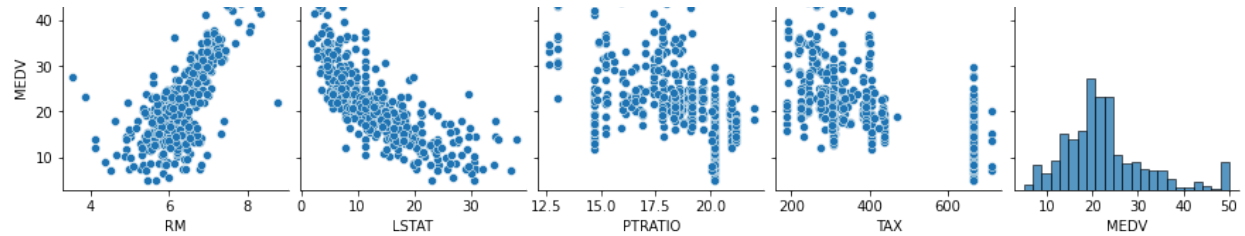
	RM	LSTAT	PTRATIO	TAX	MEDV
0	6.575	4.98	15.3	296	24.0
1	6.421	9.14	17.8	242	21.6
2	7.185	4.03	17.8	242	34.7
3	6.998	2.94	18.7	222	33.4
4	7.147	11.43	18.7	222	36.2
...	...	...	...	...	...
501	6.593	11.43	21.0	273	22.4
502	6.120	9.08	21.0	273	20.6
503	6.976	5.64	21.0	273	23.9
504	6.794	6.48	21.0	273	22.0
505	6.030	7.88	21.0	273	11.9

506 rows × 5 columns

In [50]:
```
sns.pairplot(df1)
```

Out[50]:
```
<seaborn.axisgrid.PairGrid at 0x7df960b58490>
```





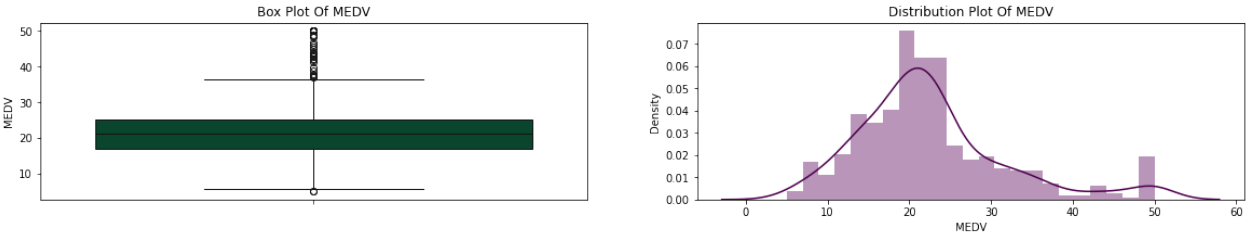
In [58]:
```
d=df1.describe()
```

In [59]:
```
plt.figure(figsize=(20,3))

plt.subplot(1,2,1)
sns.boxplot(df1.MEDV,color='#005030')
plt.title('Box Plot Of MEDV')

plt.subplot(1,2,2)
sns.distplot(a=df1.MEDV,color='#500050')
plt.title('Distribution Plot Of MEDV')

plt.show()
```



In [61]:
```
MEDV_Q3=d['MEDV']['75%']
MEDV_Q3
```

Out[61]: 25.0

In [62]:
```
MEDV_Q1=d['MEDV']['25%']
MEDV_Q1
```

Out[62]: 17.025

In [64]:
```
MEDV_IQR=MEDV_Q3-MEDV_Q1
MEDV_IQR
```

Out[64]: 7.975000000000001

In [72]:
```
MEDV_UV=MEDV_Q3+1.25*MEDV_IQR
MEDV_UV
```

Out[72]: 34.96875

In [73]:
```
MEDV_NV=MEDV_Q1-1.25*MEDV_IQR
MEDV_NV
```

Out[73]: 7.056249999999997

In [74]:
```
df1[df1['MEDV']>MEDV_UV].sort_values(by=['MEDV','RM'])
```

Out[74]:

	RM	LSTAT	PTRATIO	TAX	MEDV
279	6.812	4.85	14.9	216	35.1
273	7.691	6.58	18.6	223	35.2
281	6.968	4.59	14.9	216	35.4
55	7.249	4.81	17.9	226	35.4
258	7.333	7.79	13.0	264	36.0
304	7.236	6.93	18.4	222	36.1
181	6.144	9.45	17.8	193	36.2
4	7.147	11.43	18.7	222	36.2
192	7.178	2.87	15.2	398	36.4
264	7.206	8.10	13.0	264	36.5
190	6.951	5.10	15.2	398	37.0
179	6.980	5.04	17.8	193	37.2
291	7.148	3.56	19.2	245	37.3
226	8.040	11.43	17.4	307	37.6
182	7.155	4.82	17.8	193	37.9
97	8.069	4.21	18.0	276	38.7
180	7.765	7.56	17.8	193	39.8
157	6.943	4.59	14.7	403	41.3
232	8.337	2.47	17.4	307	41.7
202	7.610	3.11	14.7	348	42.3
253	8.259	3.54	19.1	330	42.8
261	7.520	7.26	13.0	264	43.1
268	7.470	3.16	13.0	264	43.5
98	7.820	3.57	18.0	276	43.8
256	7.454	3.11	15.9	244	44.0

	RM	LSTAT	PTRATIO	TAX	MEDV
224	8.266	4.14	17.4	307	44.8
280	7.820	3.76	14.9	216	45.4
282	7.645	3.01	14.9	216	46.0
228	7.686	11.43	17.4	307	46.7
233	8.247	3.95	17.4	307	48.3
203	7.853	3.81	14.7	224	48.5
262	8.398	5.91	13.0	264	48.8
368	4.970	3.26	20.2	666	50.0
372	5.875	8.88	20.2	666	50.0
371	6.216	9.53	20.2	666	50.0
369	6.683	3.73	20.2	666	50.0
370	7.016	2.96	20.2	666	50.0
161	7.489	1.73	14.7	403	50.0
162	7.802	1.92	14.7	403	50.0
186	7.831	4.45	17.8	193	50.0
195	7.875	2.97	14.4	255	50.0
283	7.923	3.16	13.6	198	50.0
166	7.929	3.70	14.7	403	50.0
204	8.034	2.88	14.7	224	50.0
267	8.297	7.44	13.0	264	50.0
163	8.375	3.32	14.7	403	50.0
257	8.704	5.12	13.0	264	50.0
225	8.725	4.63	17.4	307	50.0

In [78]:
```
print(f'Shape of thae dataset before removing outliers:{df1.shape}')
df2=df1[~(df1['MEDV']==50)]
print(f'Shape of thae dataset after removing outliers:{df2.shape}')
```

```
Shape of thae dataset before removing outliers:(506, 5)
Shape of thae dataset after removing outliers:(490, 5)
```

In [ ]:
```

```