EDA_HW_boston_CMPE188

February 9, 2025

Worked with: - Trevor Mathisen - Viet Nguyen

[22]:
```
import pandas as pd
import matplotlib.pyplot as plt
from pandas import set_option
from pandas import read_csv
from sklearn.preprocessing import StandardScaler
from sklearn.preprocessing import Normalizer
from numpy import set_printoptions
import seaborn as sns
from pandas.plotting import scatter_matrix
```

1. In this assignment you will perform exploratory data analysis on the Boston dataset.
2. The dataset has been provided on Canvas.
3. Load the dataset into a Pandas dataframe.
4. Clean the data (if needed).

[23]:
```
filename = 'boston.csv'
data = read_csv(filename)
set_printoptions(precision=3)
data = data.drop('index', axis=1)
data.head(5)
```

[23]:
```
      crim    zn  indus  chas  nox   rm   age  dis  rad  tax  ptratio  black
lstat  medv
0  6.3e-03  18.0    2.3     0  0.5  6.6  65.2  4.1    1  296     15.3  396.9
5.0  24.0
1  2.7e-02   0.0    7.1     0  0.5  6.4  78.9  5.0    2  242     17.8  396.9
9.1  21.6
2  2.7e-02   0.0    7.1     0  0.5  7.2  61.1  5.0    2  242     17.8  392.8
4.0  34.7
3  3.2e-02   0.0    2.2     0  0.5  7.0  45.8  6.1    3  222     18.7  394.6
2.9  33.4
4  6.9e-02   0.0    2.2     0  0.5  7.1  54.2  6.1    3  222     18.7  396.9
5.3  36.2
```

[24]:
```
# Check for missing values
print(data.isnull().sum())
```

```
crim       0
zn         0
indus      0
chas       0
nox        0
rm         0
age        0
dis        0
rad        0
tax        0
ptratio    0
black      0
lstat      0
medv       0
dtype: int64
```

Nothing to clean

5. The output in this data set is Medv (median price). The rest of the columns are considered input. Separate the data into an input and output dataframes/Series. You can ignore/eliminate categorical data.

[25]:
```python
# Split into input/output datasets (medv is output)
array = data.values
Y1 = data['medv']
X1 = data.drop('medv', axis=1)
X1names = X1.columns
```

6. Perform normalization and standardization on the data. We normally normalize and standardize the input frame and keep the output intact.
7. Put the new normalized input data frame and the output into a new data frame called data_norm. Do the same for standardized data. Call the new data frame for standardized data as data_stand.

[26]:
```python
data_norm = X1.copy()
# Normalize
norm_scaler = Normalizer().fit(data_norm)
data_norm = norm_scaler.transform(data_norm)
# add output to normalized data
data_norm = pd.DataFrame(data_norm, columns=X1names)
data_norm['medv'] = Y1

data_stand = X1.copy()
# Standardize
stand_scaler = StandardScaler().fit(data_stand)
data_stand = stand_scaler.transform(data_stand)
# add output to standardized data
data_stand = pd.DataFrame(data_stand, columns=X1names)
data_stand['medv'] = Y1
```

```
data_objects = ((data_norm, 'data_norm'), (data_stand, 'data_stand'), (data,␣
↪"data_raw"))
```

8. Perform basic EDA, i.e. descriptive stats, plot the histograms and match/verify with descriptive stats.

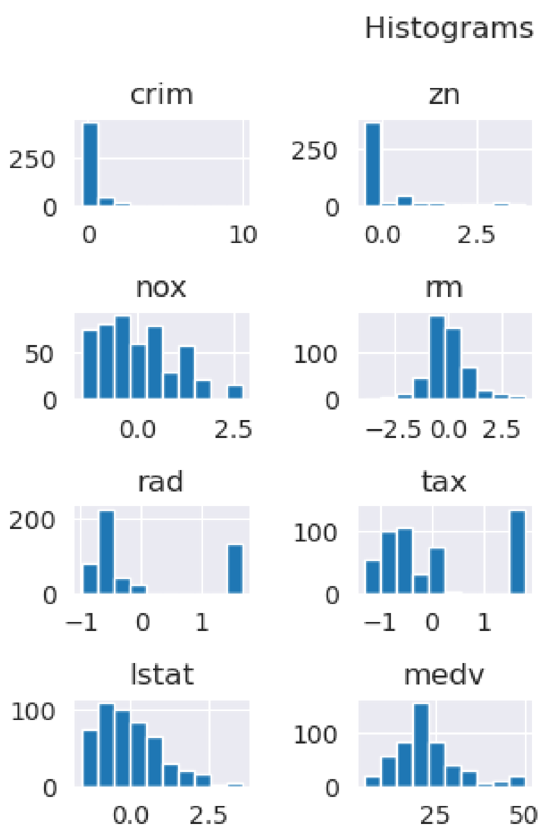
[27]:
```
# Descriptive stats
set_option('display.width', 100)
set_option('display.precision', 1)
for data, name in data_objects:
    print(f"Data: {name}")
    print(data.describe())
```

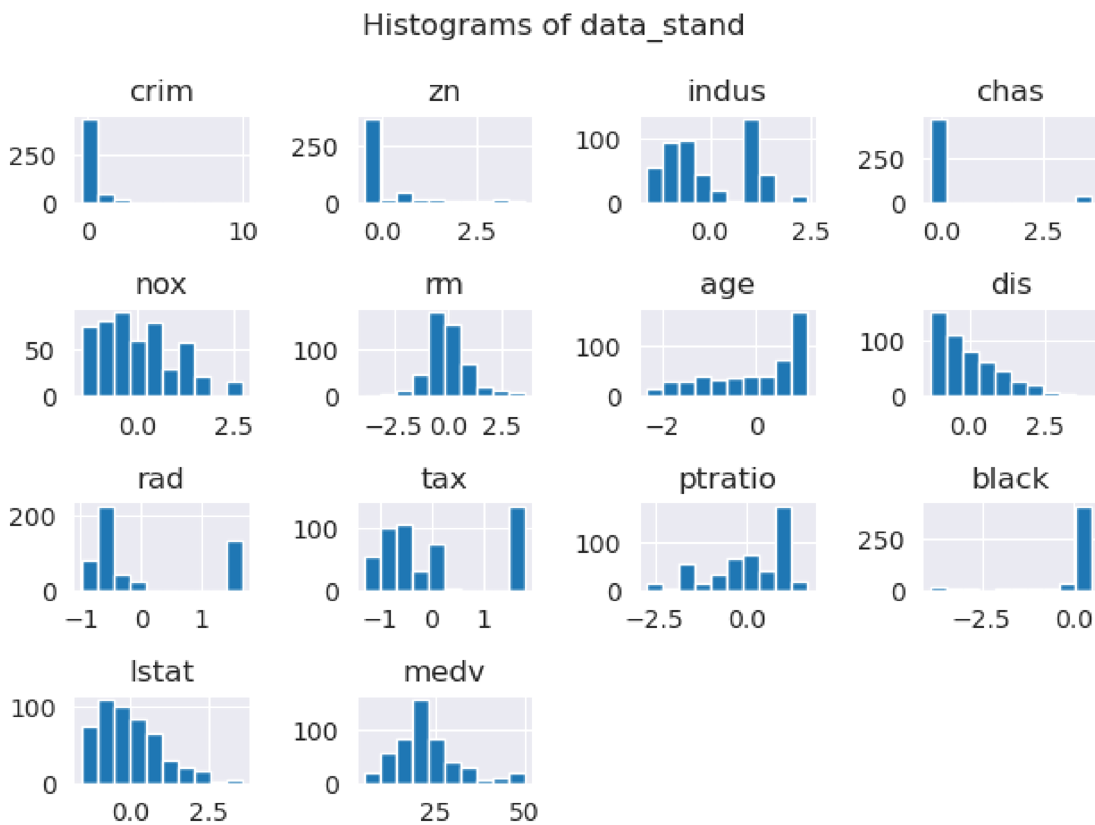
```
Data: data_norm
          crim        zn     indus      chas       nox        rm       age       dis
rad     tax  \
count  5.1e+02  5.1e+02  5.1e+02  5.1e+02  5.1e+02  5.1e+02  5.1e+02  5.1e+02
5.1e+02  506.0
mean   5.1e-03  2.3e-02  1.9e-02  1.3e-04  9.9e-04  1.2e-02  1.2e-01  7.3e-03
1.5e-02    0.7
std    1.2e-02  4.7e-02  1.1e-02  4.8e-04  1.9e-04  2.7e-03  4.8e-02  4.6e-03
1.1e-02    0.2
min    1.3e-05  0.0e+00  9.6e-04  0.0e+00  6.6e-04  4.7e-03  6.4e-03  1.5e-03
1.7e-03    0.4
25%    1.6e-04  0.0e+00  1.1e-02  0.0e+00  8.7e-04  9.4e-03  8.6e-02  3.3e-03
7.9e-03    0.6
50%    5.0e-04  0.0e+00  1.7e-02  0.0e+00  9.7e-04  1.2e-02  1.3e-01  6.4e-03
1.0e-02    0.7
75%    5.2e-03  2.5e-02  2.4e-02  0.0e+00  1.1e-03  1.3e-02  1.6e-01  1.1e-02
3.1e-02    0.9
max    1.1e-01  2.1e-01  6.1e-02  2.4e-03  2.1e-03  1.8e-02  2.4e-01  2.8e-02
3.6e-02    1.0

       ptratio    black    lstat   medv
count  5.1e+02  5.1e+02  5.1e+02  506.0
mean   3.3e-02  6.6e-01  2.2e-02   22.5
std    6.0e-03  2.1e-01  1.1e-02    9.2
min    2.4e-02  4.8e-04  3.1e-03    5.0
25%    2.7e-02  5.1e-01  1.4e-02   17.0
50%    3.3e-02  7.4e-01  2.0e-02   21.2
75%    3.8e-02  8.1e-01  2.8e-02   25.0
max    5.3e-02  8.9e-01  7.0e-02   50.0
Data: data_stand
          crim        zn     indus      chas       nox        rm       age       dis
rad     tax  \
count  5.1e+02  5.1e+02  5.1e+02  5.1e+02  5.1e+02  5.1e+02  5.1e+02  5.1e+02
5.1e+02  506.0
mean  -1.1e-16  7.9e-17  2.1e-16 -3.5e-17 -2.0e-16 -1.1e-16 -1.5e-16 -8.4e-17
```

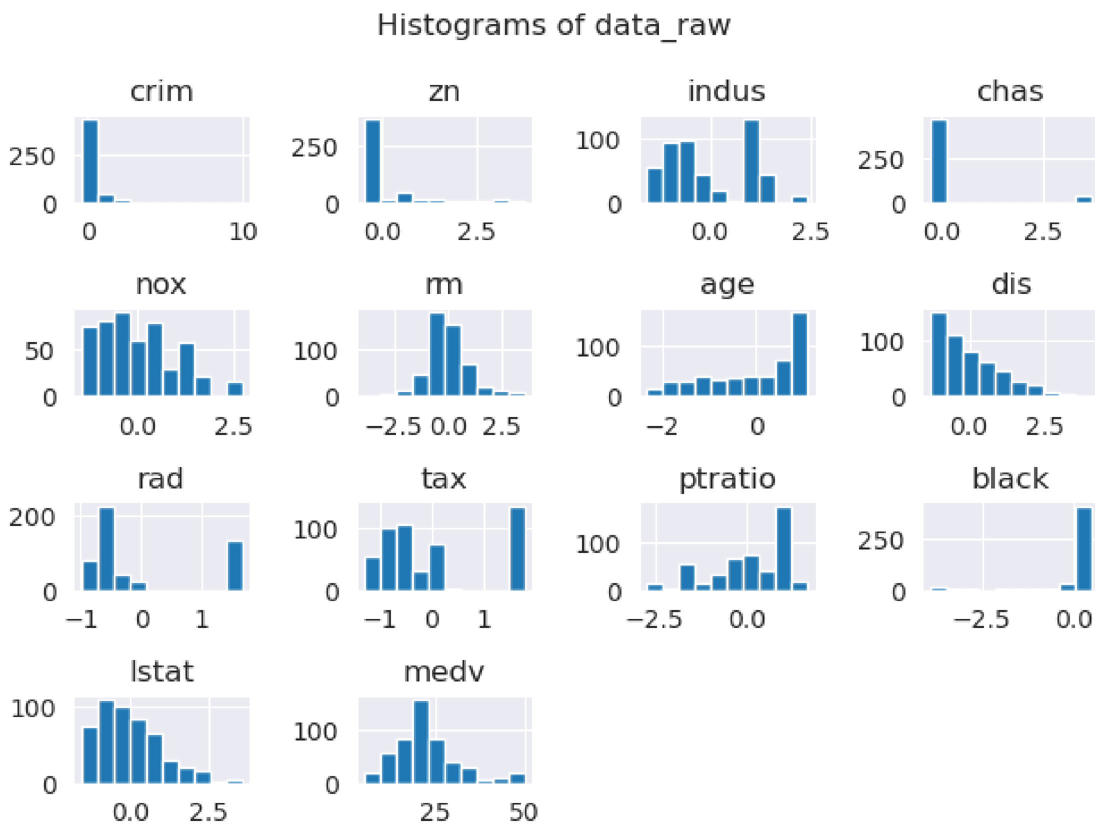
```
-1.1e-16    0.0
std    1.0e+00  1.0e+00  1.0e+00  1.0e+00  1.0e+00  1.0e+00  1.0e+00  1.0e+00
1.0e+00    1.0
min   -4.2e-01 -4.9e-01 -1.6e+00 -2.7e-01 -1.5e+00 -3.9e+00 -2.3e+00 -1.3e+00
-9.8e-01   -1.3
25%   -4.1e-01 -4.9e-01 -8.7e-01 -2.7e-01 -9.1e-01 -5.7e-01 -8.4e-01 -8.1e-01
-6.4e-01   -0.8
50%   -3.9e-01 -4.9e-01 -2.1e-01 -2.7e-01 -1.4e-01 -1.1e-01  3.2e-01 -2.8e-01
-5.2e-01   -0.5
75%    7.4e-03  4.9e-02  1.0e+00 -2.7e-01  6.0e-01  4.8e-01  9.1e-01  6.6e-01
1.7e+00    1.5
max    9.9e+00  3.8e+00  2.4e+00  3.7e+00  2.7e+00  3.6e+00  1.1e+00  4.0e+00
1.7e+00    1.8

       ptratio    black    lstat   medv
count  5.1e+02  5.1e+02  5.1e+02  506.0
mean  -4.2e-16 -7.4e-16 -3.1e-16   22.5
std    1.0e+00  1.0e+00  1.0e+00    9.2
min   -2.7e+00 -3.9e+00 -1.5e+00    5.0
25%   -4.9e-01  2.1e-01 -8.0e-01   17.0
50%    2.7e-01  3.8e-01 -1.8e-01   21.2
75%    8.1e-01  4.3e-01  6.0e-01   25.0
max    1.6e+00  4.4e-01  3.5e+00   50.0
Data: data_raw
          crim     zn  indus     chas    nox     rm    age    dis    rad    tax
ptratio  black  \
count  5.1e+02  506.0  506.0  5.1e+02  506.0  506.0  506.0  506.0  506.0  506.0
506.0  506.0
mean   3.6e+00   11.4   11.1  6.9e-02    0.6    6.3   68.6    3.8    9.5  408.2
18.5  356.7
std    8.6e+00   23.3    6.9  2.5e-01    0.1    0.7   28.1    2.1    8.7  168.5
2.2   91.3
min    6.3e-03    0.0    0.5  0.0e+00    0.4    3.6    2.9    1.1    1.0  187.0
12.6    0.3
25%    8.2e-02    0.0    5.2  0.0e+00    0.4    5.9   45.0    2.1    4.0  279.0
17.4  375.4
50%    2.6e-01    0.0    9.7  0.0e+00    0.5    6.2   77.5    3.2    5.0  330.0
19.1  391.4
75%    3.7e+00   12.5   18.1  0.0e+00    0.6    6.6   94.1    5.2   24.0  666.0
20.2  396.2
max    8.9e+01  100.0   27.7  1.0e+00    0.9    8.8  100.0   12.1   24.0  711.0
22.0  396.9

       lstat   medv
count  506.0  506.0
mean    12.7   22.5
std      7.1    9.2
min      1.7    5.0
```

```
25%      6.9   17.0
50%     11.4   21.2
75%     17.0   25.0
max     38.0   50.0
```

[28]:
```python
# Histograms
for data, name in data_objects:
    data_stand.hist()
    plt.suptitle(f"Histograms of {name}")
    plt.tight_layout()
    plt.show()
```




Histograms of data_stand
crim
zn
indus
chas
nox
rm
age
dis
rad
tax
ptratio
black
lstat
medv

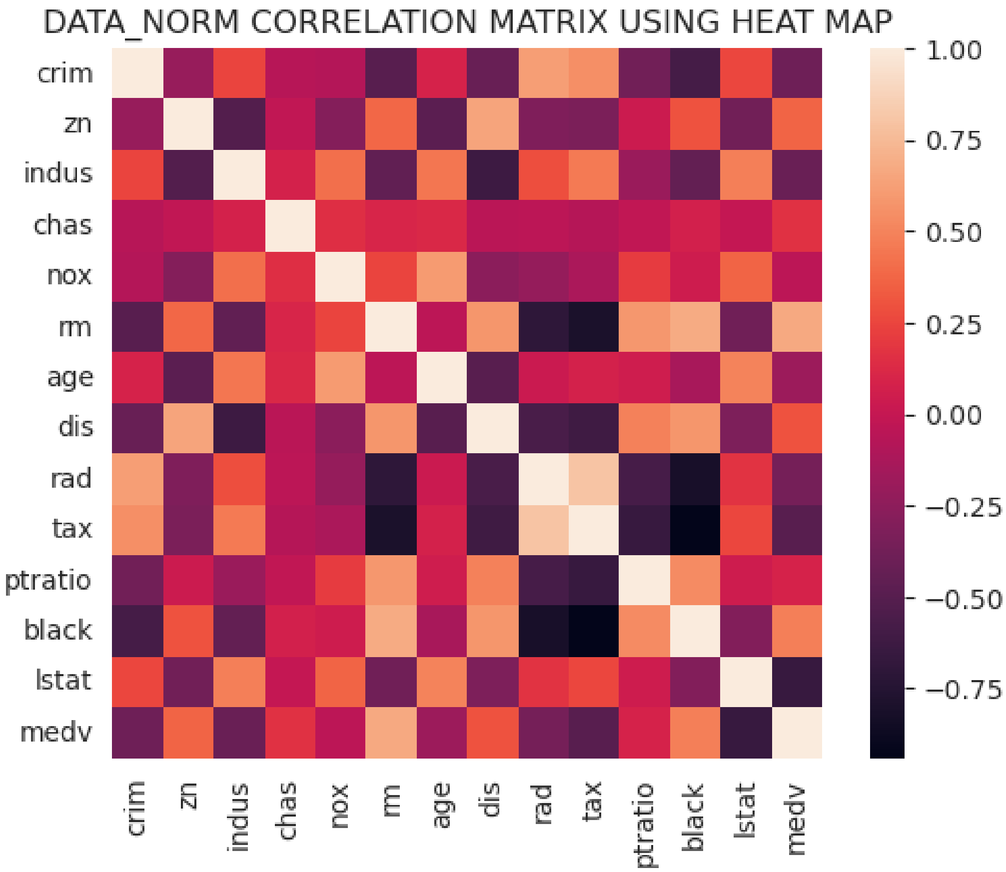


9. Continue with correlation analysis (calculate correlation and plot correlation heatmap) and scatter plots.

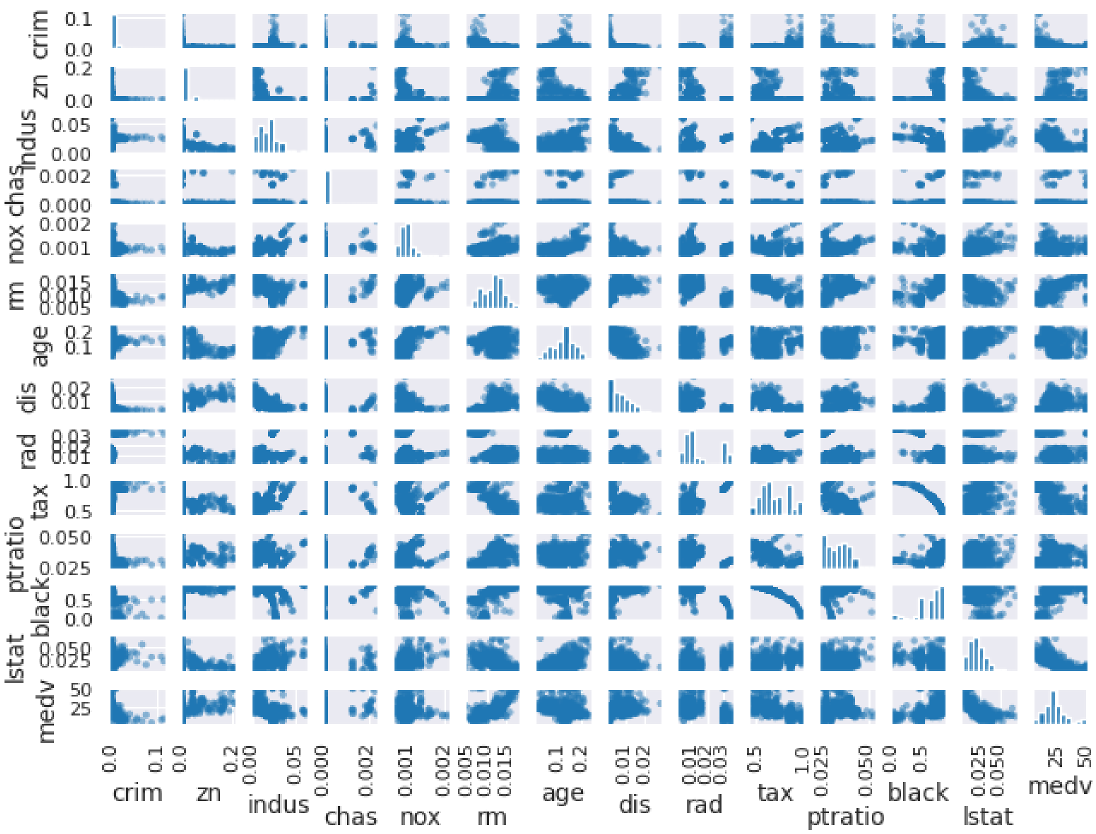
[34]:
```
for data, name in data_objects:
    plt.figure() # new plot
    #plt.tight_layout()
    corMat = data_norm.corr(method='pearson')
    print(corMat)
    ## plot correlation matrix as a heat map
    sns.heatmap(corMat, square=True)
    plt.yticks(rotation=0)
    plt.xticks(rotation=90)
    plt.title(f"{name.upper()} CORRELATION MATRIX USING HEAT MAP")
    plt.show()

    ## scatter plot of all data
    plt.figure()
    # # The output overlaps itself, resize it to display better (w padding)
    scatter_matrix(data_norm)
    plt.tight_layout(pad=0.1)
    plt.show()
```

	crim	zn	indus	chas	nox	rm	age	dis	rad	tax
crim	1.0e+00	-2.1e-01	2.5e-01	-6.3e-02	-8.2e-02	-4.9e-01	8.7e-02	-4.2e-01	6.1e-01	5.5e-01
zn	-2.1e-01	1.0e+00	-5.2e-01	-1.9e-02	-3.0e-01	3.9e-01	-4.8e-01	6.4e-01	-3.1e-01	-3.3e-01
indus	2.5e-01	-5.2e-01	1.0e+00	8.2e-02	4.1e-01	-4.5e-01	4.5e-01	-6.3e-01	2.8e-01	4.6e-01
chas	-6.3e-02	-1.9e-02	8.2e-02	1.0e+00	1.5e-01	1.0e-01	1.2e-01	-5.2e-02	-4.0e-02	-7.5e-02
nox	-8.2e-02	-3.0e-01	4.1e-01	1.5e-01	1.0e+00	2.5e-01	6.0e-01	-2.6e-01	-2.1e-01	-1.2e-01
rm	-4.9e-01	3.9e-01	-4.5e-01	1.0e-01	2.5e-01	1.0e+00	-3.9e-02	5.8e-01	-7.0e-01	-8.0e-01
age	8.7e-02	-4.8e-01	4.5e-01	1.2e-01	6.0e-01	-3.9e-02	1.0e+00	-4.9e-01	2.5e-02	7.9e-02
dis	-4.2e-01	6.4e-01	-6.3e-01	-5.2e-02	-2.6e-01	5.8e-01	-4.9e-01	1.0e+00	-5.6e-01	-6.2e-01
rad	6.1e-01	-3.1e-01	2.8e-01	-4.0e-02	-2.1e-01	-7.0e-01	2.5e-02	-5.6e-01	1.0e+00	8.0e-01
tax	5.5e-01	-3.3e-01	4.6e-01	-7.5e-02	-1.2e-01	-8.0e-01	7.9e-02	-6.2e-01	8.0e-01	1.0e+00
ptratio	-3.8e-01	3.3e-02	-2.0e-01	-1.6e-02	2.1e-01	5.9e-01	4.8e-02	4.9e-01	-5.8e-01	-6.5e-01
black	-5.9e-01	3.0e-01	-4.5e-01	7.0e-02	4.2e-02	6.8e-01	-1.3e-01	5.8e-01	-8.1e-01	-9.4e-01
lstat	2.5e-01	-3.7e-01	4.8e-01	-2.4e-03	3.7e-01	-3.8e-01	5.0e-01	-3.3e-01	1.7e-01	2.6e-01
medv	-3.9e-01	3.7e-01	-4.1e-01	1.7e-01	-4.5e-02	6.7e-01	-1.9e-01	3.0e-01	-3.6e-01	-5.0e-01

	ptratio	black	lstat	medv
crim	-3.8e-01	-5.9e-01	2.5e-01	-3.9e-01
zn	3.3e-02	3.0e-01	-3.7e-01	3.7e-01
indus	-2.0e-01	-4.5e-01	4.8e-01	-4.1e-01
chas	-1.6e-02	7.0e-02	-2.4e-03	1.7e-01
nox	2.1e-01	4.2e-02	3.7e-01	-4.5e-02
rm	5.9e-01	6.8e-01	-3.8e-01	6.7e-01
age	4.8e-02	-1.3e-01	5.0e-01	-1.9e-01
dis	4.9e-01	5.8e-01	-3.3e-01	3.0e-01
rad	-5.8e-01	-8.1e-01	1.7e-01	-3.6e-01
tax	-6.5e-01	-9.4e-01	2.6e-01	-5.0e-01
ptratio	1.0e+00	5.4e-01	4.3e-02	8.6e-02
black	5.4e-01	1.0e+00	-3.0e-01	4.8e-01
lstat	4.3e-02	-3.0e-01	1.0e+00	-6.5e-01
medv	8.6e-02	4.8e-01	-6.5e-01	1.0e+00

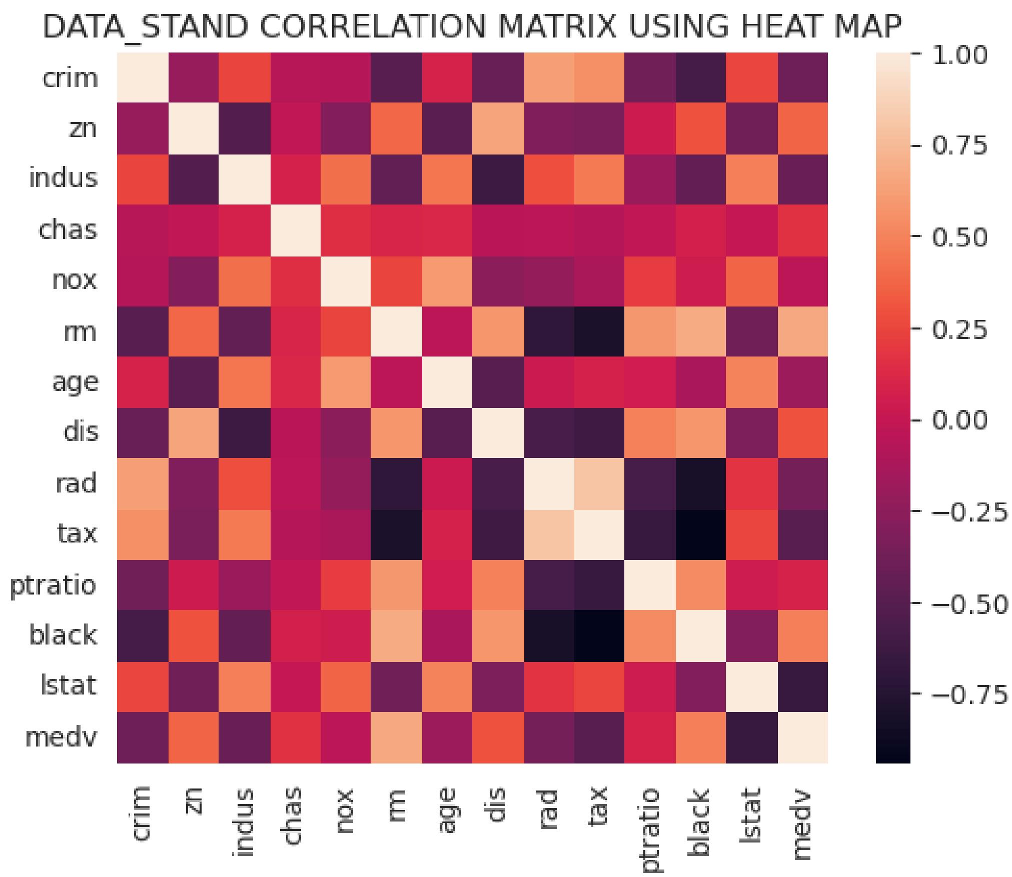



```
<Figure size 640x480 with 0 Axes>
```

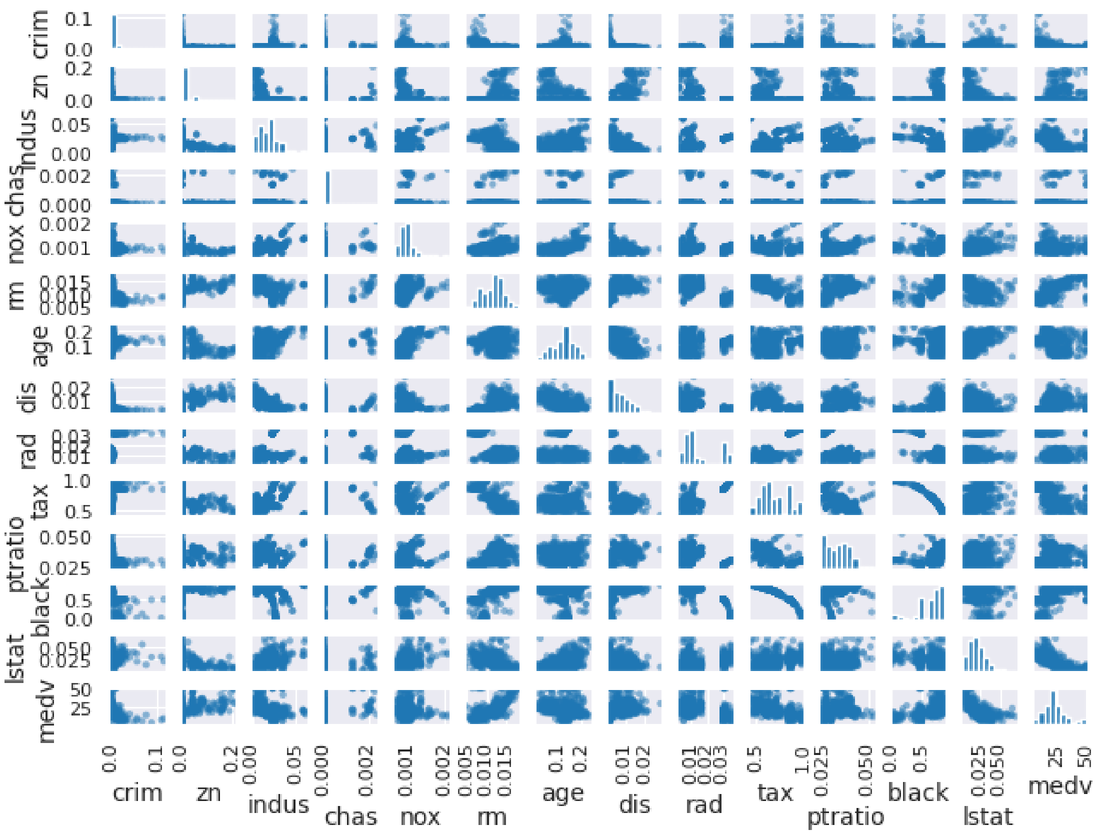



```
            crim        zn     indus      chas       nox        rm       age       dis
rad       tax  \
crim     1.0e+00 -2.1e-01  2.5e-01 -6.3e-02 -8.2e-02 -4.9e-01  8.7e-02 -4.2e-01
6.1e-01  5.5e-01
zn      -2.1e-01  1.0e+00 -5.2e-01 -1.9e-02 -3.0e-01  3.9e-01 -4.8e-01  6.4e-01
-3.1e-01 -3.3e-01
indus    2.5e-01 -5.2e-01  1.0e+00  8.2e-02  4.1e-01 -4.5e-01  4.5e-01 -6.3e-01
2.8e-01  4.6e-01
chas    -6.3e-02 -1.9e-02  8.2e-02  1.0e+00  1.5e-01  1.0e-01  1.2e-01 -5.2e-02
-4.0e-02 -7.5e-02
nox     -8.2e-02 -3.0e-01  4.1e-01  1.5e-01  1.0e+00  2.5e-01  6.0e-01 -2.6e-01
-2.1e-01 -1.2e-01
rm      -4.9e-01  3.9e-01 -4.5e-01  1.0e-01  2.5e-01  1.0e+00 -3.9e-02  5.8e-01
-7.0e-01 -8.0e-01
age      8.7e-02 -4.8e-01  4.5e-01  1.2e-01  6.0e-01 -3.9e-02  1.0e+00 -4.9e-01
2.5e-02  7.9e-02
dis     -4.2e-01  6.4e-01 -6.3e-01 -5.2e-02 -2.6e-01  5.8e-01 -4.9e-01  1.0e+00
-5.6e-01 -6.2e-01
rad      6.1e-01 -3.1e-01  2.8e-01 -4.0e-02 -2.1e-01 -7.0e-01  2.5e-02 -5.6e-01
1.0e+00  8.0e-01
tax      5.5e-01 -3.3e-01  4.6e-01 -7.5e-02 -1.2e-01 -8.0e-01  7.9e-02 -6.2e-01
```

```
8.0e-01  1.0e+00
ptratio -3.8e-01  3.3e-02 -2.0e-01 -1.6e-02  2.1e-01  5.9e-01  4.8e-02  4.9e-01
-5.8e-01 -6.5e-01
black   -5.9e-01  3.0e-01 -4.5e-01  7.0e-02  4.2e-02  6.8e-01 -1.3e-01  5.8e-01
-8.1e-01 -9.4e-01
lstat    2.5e-01 -3.7e-01  4.8e-01 -2.4e-03  3.7e-01 -3.8e-01  5.0e-01 -3.3e-01
1.7e-01  2.6e-01
medv    -3.9e-01  3.7e-01 -4.1e-01  1.7e-01 -4.5e-02  6.7e-01 -1.9e-01  3.0e-01
-3.6e-01 -5.0e-01

         ptratio    black    lstat     medv
crim    -3.8e-01 -5.9e-01  2.5e-01 -3.9e-01
zn       3.3e-02  3.0e-01 -3.7e-01  3.7e-01
indus   -2.0e-01 -4.5e-01  4.8e-01 -4.1e-01
chas    -1.6e-02  7.0e-02 -2.4e-03  1.7e-01
nox      2.1e-01  4.2e-02  3.7e-01 -4.5e-02
rm       5.9e-01  6.8e-01 -3.8e-01  6.7e-01
age      4.8e-02 -1.3e-01  5.0e-01 -1.9e-01
dis      4.9e-01  5.8e-01 -3.3e-01  3.0e-01
rad     -5.8e-01 -8.1e-01  1.7e-01 -3.6e-01
tax     -6.5e-01 -9.4e-01  2.6e-01 -5.0e-01
ptratio  1.0e+00  5.4e-01  4.3e-02  8.6e-02
black    5.4e-01  1.0e+00 -3.0e-01  4.8e-01
lstat    4.3e-02 -3.0e-01  1.0e+00 -6.5e-01
medv     8.6e-02  4.8e-01 -6.5e-01  1.0e+00
```

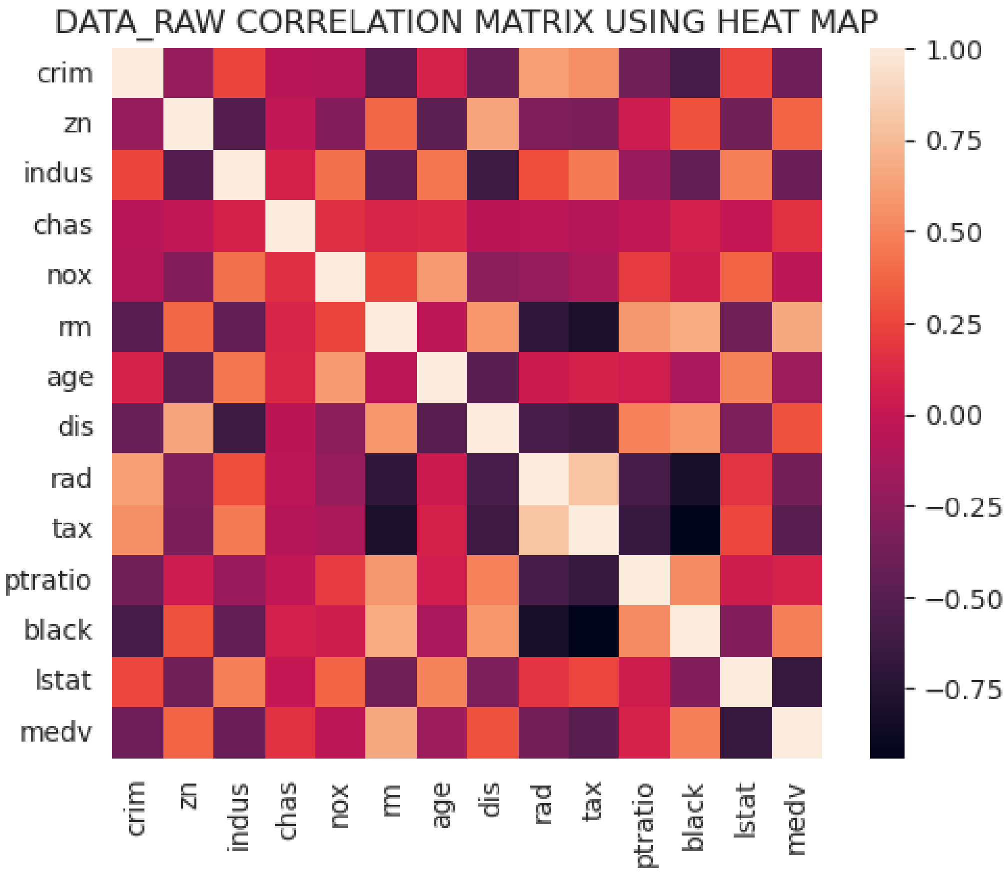



```
<Figure size 640x480 with 0 Axes>
```

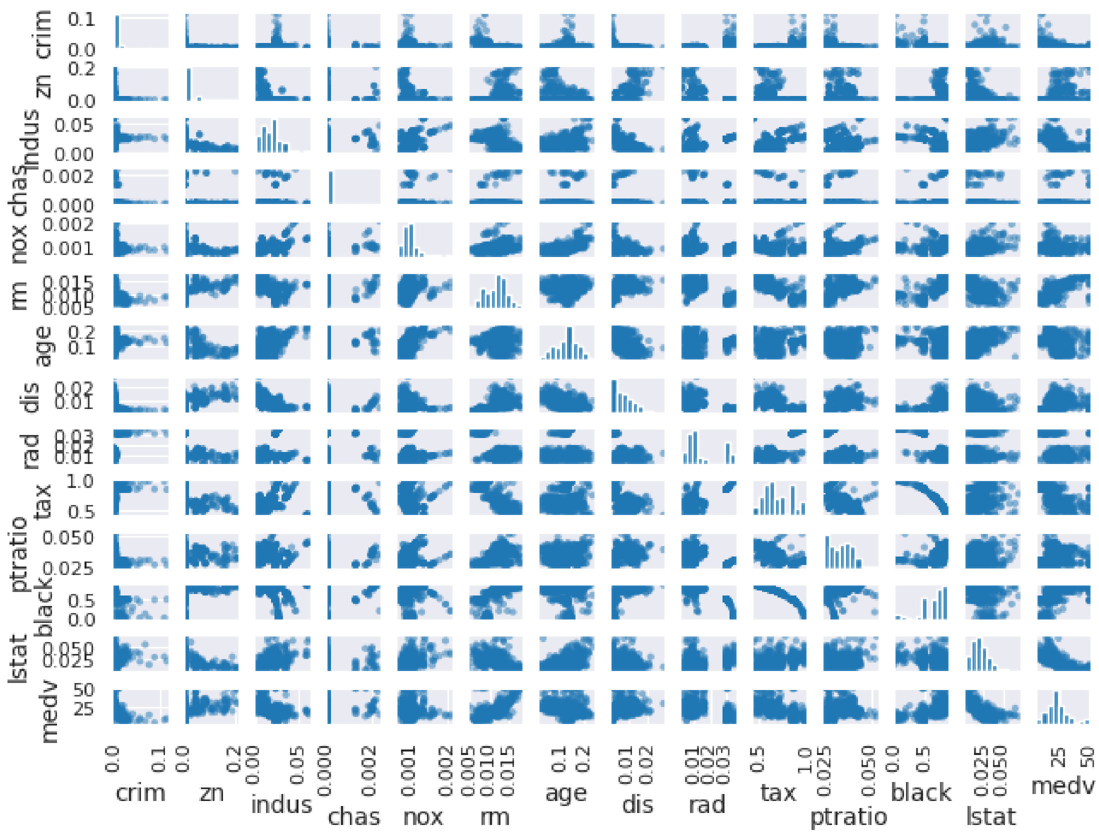



```
            crim        zn     indus      chas       nox        rm       age       dis  \
rad       tax  \
crim     1.0e+00 -2.1e-01  2.5e-01 -6.3e-02 -8.2e-02 -4.9e-01  8.7e-02 -4.2e-01
6.1e-01  5.5e-01
zn      -2.1e-01  1.0e+00 -5.2e-01 -1.9e-02 -3.0e-01  3.9e-01 -4.8e-01  6.4e-01
-3.1e-01 -3.3e-01
indus    2.5e-01 -5.2e-01  1.0e+00  8.2e-02  4.1e-01 -4.5e-01  4.5e-01 -6.3e-01
2.8e-01  4.6e-01
chas    -6.3e-02 -1.9e-02  8.2e-02  1.0e+00  1.5e-01  1.0e-01  1.2e-01 -5.2e-02
-4.0e-02 -7.5e-02
nox     -8.2e-02 -3.0e-01  4.1e-01  1.5e-01  1.0e+00  2.5e-01  6.0e-01 -2.6e-01
-2.1e-01 -1.2e-01
rm      -4.9e-01  3.9e-01 -4.5e-01  1.0e-01  2.5e-01  1.0e+00 -3.9e-02  5.8e-01
-7.0e-01 -8.0e-01
age      8.7e-02 -4.8e-01  4.5e-01  1.2e-01  6.0e-01 -3.9e-02  1.0e+00 -4.9e-01
2.5e-02  7.9e-02
dis     -4.2e-01  6.4e-01 -6.3e-01 -5.2e-02 -2.6e-01  5.8e-01 -4.9e-01  1.0e+00
-5.6e-01 -6.2e-01
rad      6.1e-01 -3.1e-01  2.8e-01 -4.0e-02 -2.1e-01 -7.0e-01  2.5e-02 -5.6e-01
1.0e+00  8.0e-01
tax      5.5e-01 -3.3e-01  4.6e-01 -7.5e-02 -1.2e-01 -8.0e-01  7.9e-02 -6.2e-01
```

```
8.0e-01  1.0e+00
ptratio -3.8e-01  3.3e-02 -2.0e-01 -1.6e-02  2.1e-01  5.9e-01  4.8e-02  4.9e-01
-5.8e-01 -6.5e-01
black   -5.9e-01  3.0e-01 -4.5e-01  7.0e-02  4.2e-02  6.8e-01 -1.3e-01  5.8e-01
-8.1e-01 -9.4e-01
lstat    2.5e-01 -3.7e-01  4.8e-01 -2.4e-03  3.7e-01 -3.8e-01  5.0e-01 -3.3e-01
1.7e-01  2.6e-01
medv    -3.9e-01  3.7e-01 -4.1e-01  1.7e-01 -4.5e-02  6.7e-01 -1.9e-01  3.0e-01
-3.6e-01 -5.0e-01

         ptratio    black    lstat     medv
crim    -3.8e-01 -5.9e-01  2.5e-01 -3.9e-01
zn       3.3e-02  3.0e-01 -3.7e-01  3.7e-01
indus   -2.0e-01 -4.5e-01  4.8e-01 -4.1e-01
chas    -1.6e-02  7.0e-02 -2.4e-03  1.7e-01
nox      2.1e-01  4.2e-02  3.7e-01 -4.5e-02
rm       5.9e-01  6.8e-01 -3.8e-01  6.7e-01
age      4.8e-02 -1.3e-01  5.0e-01 -1.9e-01
dis      4.9e-01  5.8e-01 -3.3e-01  3.0e-01
rad     -5.8e-01 -8.1e-01  1.7e-01 -3.6e-01
tax     -6.5e-01 -9.4e-01  2.6e-01 -5.0e-01
ptratio  1.0e+00  5.4e-01  4.3e-02  8.6e-02
black    5.4e-01  1.0e+00 -3.0e-01  4.8e-01
lstat    4.3e-02 -3.0e-01  1.0e+00 -6.5e-01
medv     8.6e-02  4.8e-01 -6.5e-01  1.0e+00
```




```
<Figure size 640x480 with 0 Axes>
```



11. Identify the high correlation columns from the headmap and compare the results from those of the scatter plots. Do the results match? Explain.

The high correlation columns from the heatmap match the results from the scatter plots. However, there are deviations in the data due to outliers or clusters that affect the linearity of the graph. A negative correlation between distance and age, shown on the heat map, is reflected on the scatterplot by a negative linear slope. Rad (index of accessibility to radial highways) and tax (property tax rate) have a strong positive correlation on the heatmap, suggesting that areas with high accessibility to highways tend to have higher property tax rates. Their scatter plot has a kind of positively correlated slope with a gap in the distribution, likely caused by outliers.