In [1]:
```
# Import necessary libraries
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [14]:
```
# Task 1: Read the dataset and report the number of instances and features
df = pd.read_csv("C:\\Users\\user\\Desktop\\housing.header.txt")
```

In [15]:
```
df.shape
```

Out[15]:
```
(506, 14)
```

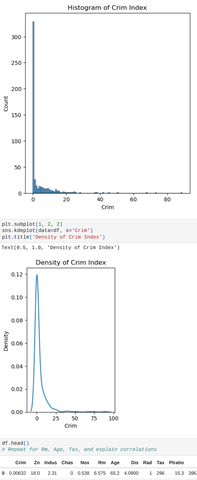
In [16]:
```
num_instances, num_features = df.shape
```

In [17]:
```
df.shape
```

Out[17]:
```
(506, 14)
```

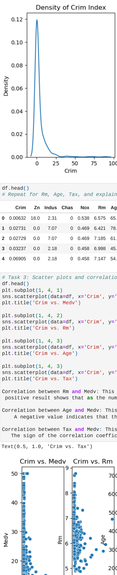
In [23]:
```
# Task 2: Show histogram and density of the Crim index
plt.figure(figsize=(12, 6))
plt.subplot(1,2,1)
sns.histplot(data=df, x='Crim', )
plt.title('Histogram of Crim Index')
```

Out[23]:
```
Text(0.5, 1.0, 'Histogram of Crim Index')
```



In [39]:
```
plt.subplot(1, 2, 2)
sns.kdeplot(data=df, x='Crim')
plt.title('Density of Crim Index')
```

Out[39]:
```
Text(0.5, 1.0, 'Density of Crim Index')
```



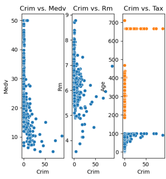
In [40]:
```
df.head()
# Repeat for Rm, Age, Tax, and explain correlations
```

Out[40]:

	Crim	Zn	Indus	Chas	Nox	Rm	Age	Dis	Rad	Tax	Ptratio	B	Lstat	Medv
0	0.00632	18.0	2.31	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	0.469	6.421	78.9	4.9671	2	242	17.8	396.90	9.14	21.6
2	0.02729	0.0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
3	0.03237	0.0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4
4	0.06905	0.0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	396.90	5.33	36.2

In [45]:
```
# Task 3: Scatter plots and correlations
df.head()
plt.subplot(1, 4, 1)
sns.scatterplot(data=df, x='Crim', y='Medv')
plt.title('Crim vs. Medv')

plt.subplot(1, 4, 2)
sns.scatterplot(data=df, x='Crim', y='Rm')
plt.title('Crim vs. Rm')

plt.subplot(1, 4, 3)
sns.scatterplot(data=df, x='Crim', y='Age')
plt.title('Crim vs. Age')

plt.subplot(1, 4, 3)
sns.scatterplot(data=df, x='Crim', y='Tax')
plt.title('Crim vs. Tax')

Correlation between Rm and Medv: This value indicates the degree to which the number of rooms (Rm) is connected with the medium house value (Medv).
  positive result shows that as the number of rooms increases, so does the house value

Correlation between Age and Medv: This value indicates how the property's age (Age) is related to the medium house value (Medv).
   A negative value indicates that the house value tends to fall as the property ages.

Correlation between Tax and Medv: This figure indicates the relationship between the property tax rate (Tax) and the median house value (Medv).
   The sign of the correlation coefficient will determine the interpretation.
```

Out[45]:
```
Text(0.5, 1.0, 'Crim vs. Tax')
```

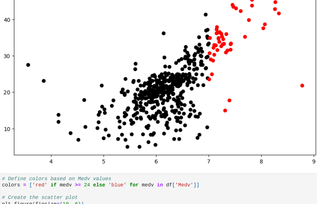


In [46]:
```
# Create the subset
subset_df = df[df['Crim'] <= 1) & (df['Rm'] > 6)]

# Display the subset
print(subset_df)
```

```
      Crim    Zn  Indus  Chas    Nox     Rm   Age     Dis  Rad  Tax  \
0    0.00632  18.0   2.31     0  0.538  6.575  65.2  4.0900    1  296
1    0.02731   0.0   7.07     0  0.469  6.421  78.9  4.9671    2  242
2    0.02729   0.0   7.07     0  0.469  7.185  61.1  4.9671    2  242
3    0.03237   0.0   2.18     0  0.458  6.998  45.8  6.0622    3  222
4    0.06905   0.0   2.18     0  0.458  7.147  54.2  6.0622    3  222
..       ...   ...    ...   ...    ...    ...   ...     ...  ...  ...
501  0.06263   0.0  11.93     0  0.573  6.593  69.1  2.4786    1  273
502  0.04527   0.0  11.93     0  0.573  6.120  76.7  2.2875    1  273
503  0.06076   0.0  11.93     0  0.573  6.976  91.0  2.1675    1  273
504  0.10959   0.0  11.93     0  0.573  6.794  89.3  2.3889    1  273
505  0.04741   0.0  11.93     0  0.573  6.030  80.8  2.5050    1  273

     Ptratio       B  Lstat  Medv
0       15.3  396.90   4.98  24.0
1       17.8  396.90   9.14  21.6
2       17.8  392.83   4.03  34.7
3       18.7  394.63   2.94  33.4
4       18.7  396.90   5.33  36.2
..       ...     ...    ...   ...
501     21.0  391.99   9.67  22.4
502     21.0  396.90   9.08  20.6
503     21.0  396.90   5.64  23.9
504     21.0  393.45   6.48  22.0
505     21.0  396.90   7.88  11.9

[236 rows x 14 columns]
```

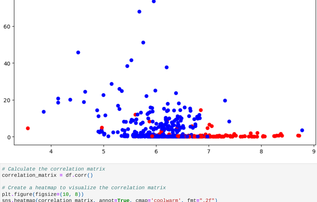
In [47]:
```
# Separate the data into two subsets based on Rm values
red_subset = df[df['Rm'] >= 7]
black_subset = df[df['Rm'] < 7]

# Create the scatter plot
plt.figure(figsize=(10, 6))
plt.scatter(black_subset['Rm'], black_subset['Medv'], color='black', label='Rm < 7')
plt.scatter(red_subset['Rm'], red_subset['Medv'], color='red', label='Rm >= 7')
```

Out[47]:
```
<matplotlib.collections.PathCollection at 0x1da5c6e45b>
```



In [52]:
```
# Define colors based on Medv values
colors = ['red' if medv >= 24 else 'blue' for medv in df['Medv']]

# Create the scatter plot
plt.figure(figsize=(10, 6))
plt.scatter(df['Rm'], df['Crim'], c=colors, )
```

Out[52]:
```
<matplotlib.collections.PathCollection at 0x13da5d36b10>
```



In [54]:
```
# Calculate the correlation matrix
correlation_matrix = df.corr()

# Create a heatmap to visualize the correlation matrix
plt.figure(figsize=(10, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Pairwise Correlation Matrix')
plt.show()


#To compute the correlation matrix for all pairs of variables in the DataFrame df, we use df.corr().

#To see the correlation matrix, we generate a heatmap with sns.heatmap().
#The annot=True option adds numerical annotations to the heatmap,
#and the color map is specified by cmap='coolwarm'.
#The fmt=".2f" parameter formats the annotations as two-decimal-place floating-point values.

#We choose a title for the heatmap.

#Finally, we display the heatmap with plt.show().
```

```
  Cell In[54], line 11
    To compute the correlation matrix for all pairs of variables in the DataFrame df, we use df.corr().
       ^
SyntaxError: invalid syntax
```

In [ ]:
```
#Please explain which variable is mostly positively correlated to Medv (medium house value),
#and which variable is mostly negatively correlated to Medv

#Here are the variables that are mostly favorably and adversely connected with "Medv" (middle house value)
#based on the correlation matrix:

#Correlations with "Medv" are mostly positive:
Answer:

#The positive correlation coefficient for "rm" (number of rooms) is roughly 0.70.
#This illustrates that as the number of rooms increases, so does the average property value.
#"Rm" has a largely good relationship with "Medv."
#Mostly Correlated with "Medv" in a Negative Way:
```

In [57]:
```
# A list of attributes to plot against Medv
attributes = ['Crim', 'Rm', 'Age', 'Tax']

# Create scatterplots for each attribute
plt.figure(figsize=(15, 10))

for i, attribute in enumerate(attributes, 1):
    plt.subplot(2, 2, i)
    sns.scatterplot(data=df, x=attribute, y='Medv', alpha=0.5)
    plt.title(f'Scatter Plot of {attribute} vs. Medv')

plt.tight_layout()
plt.show()
```

In [ ]:
```

```

In [ ]:
```

```