Activity 5: Generating Statistics from a CSV File

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

In [2]:
```
file = "boston_housing.csv"
HousingData=pd.read_csv(file)
```

In [12]:
```
HousingData.head(10)
```

Out[12]:

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT
0	0.00632	18.0	2.31	0	0.538	6.575	65.2	4.0900	1	296	15.3	396.90	4.98
1	0.02731	0.0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.90	9.14
2	0.02729	0.0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03
3	0.03237	0.0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	2.94
4	0.06905	0.0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	396.90	5.33
5	0.02985	0.0	2.18	0	0.458	6.430	58.7	6.0622	3	222	18.7	394.12	5.21
6	0.08829	12.5	7.87	0	0.524	6.012	66.6	5.5605	5	311	15.2	395.60	12.43
7	0.14455	12.5	7.87	0	0.524	6.172	96.1	5.9505	5	311	15.2	396.90	19.15
8	0.21124	12.5	7.87	0	0.524	5.631	100.0	6.0821	5	311	15.2	386.63	29.93
9	0.17004	12.5	7.87	0	0.524	6.004	85.9	6.5921	5	311	15.2	386.71	17.10

In [3]:
```
ignoreColumns = ['CHAS','NOX','B','LSTAT']
```

In [4]:
```
columns = HousingData.columns
columnList = columns.tolist()
columnList
```

Out[4]:
```
['CRIM',
 'ZN',
 'INDUS',
 'CHAS',
 'NOX',
 'RM',
 'AGE',
 'DIS',
 'RAD',
 'TAX',
 'PTRATIO',
 'B',
 'LSTAT',
 'PRICE']
```

In [5]:
```
pickColumns = [colName for colName in columnList if colName not in ignor
eColumns]
```

In [6]:
```
pickColumns
```

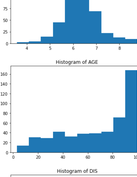
Out[6]:
```
['CRIM', 'ZN', 'INDUS', 'RM', 'AGE', 'DIS', 'RAD', 'TAX', 'PTRATIO', 'PR
ICE']
```

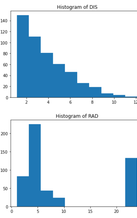
In [7]:
```
HousingData = HousingData[pickColumns]
HousingData.tail(7)
```

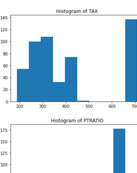
Out[7]:

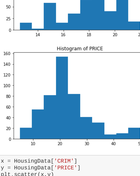
	CRIM	ZN	INDUS	RM	AGE	DIS	RAD	TAX	PTRATIO	PRICE
499	0.17783	0.0	9.69	5.569	73.5	2.3999	6	391	19.2	17.5
500	0.22438	0.0	9.69	6.027	79.7	2.4982	6	391	19.2	16.8
501	0.06263	0.0	11.93	6.593	69.1	2.4786	1	273	21.0	22.4
502	0.04527	0.0	11.93	6.120	76.7	2.2875	1	273	21.0	20.6
503	0.06076	0.0	11.93	6.976	91.0	2.1675	1	273	21.0	23.9
504	0.10959	0.0	11.93	6.794	89.3	2.3889	1	273	21.0	22.0
505	0.04741	0.0	11.93	6.030	80.8	2.5050	1	273	21.0	11.9

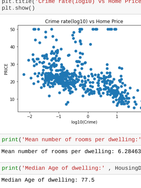
In [8]:
```
for colName in HousingData.columns:
    plt.title('Histogram of ' + colName)
    plt.hist(HousingData[colName])
    plt.show()
```

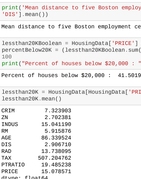


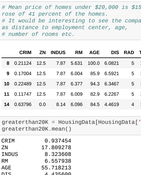


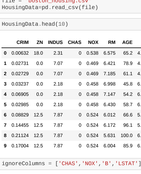




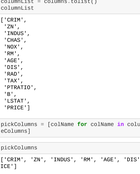








In [9]:
```
x = HousingData['CRIM']
y = HousingData['PRICE']
plt.scatter(x,y)
plt.xlabel('Crime')
plt.ylabel('PRICE')
plt.title('Crime rate vs Home Price')
plt.show()
```



In [10]:
```
x = np.log10(HousingData['CRIM'])
y = HousingData['PRICE']
plt.scatter(x,y)
plt.xlabel('log10(Crime)')
plt.ylabel('PRICE')
plt.title('Crime rate(log10) vs Home Price')
plt.show()
```

In [11]:
```
print('Mean number of rooms per dwelling:' , HousingData['RM'].mean())
```

```
Mean number of rooms per dwelling: 6.284634387351779
```

In [12]:
```
print('Median Age of dwelling:' , HousingData['AGE'].median())
```

```
Median Age of dwelling: 77.5
```

In [13]:
```
print('Mean distance to five Boston employment centers:' , HousingData[
'DIS'].mean())
```

```
Mean distance to five Boston employment centers: 3.795042687747036
```

In [14]:
```
lessthan20KBoolean = HousingData['PRICE'] < 20
percentBelow20K = (lessthan20KBoolean.sum()/lessthan20KBoolean.count())*
100
print("Percent of houses below $20,000 : ",percentBelow20K)
```

```
Percent of houses below $20,000 :  41.50197628458498
```

In [15]:
```
lessthan20K = HousingData[HousingData['PRICE'] < 20]
lessthan20K.mean()
```

Out[15]:
```
CRIM         7.323903
ZN           2.702381
INDUS       15.041190
RM           5.915876
AGE         86.339524
DIS          2.998710
RAD         13.738095
TAX        507.294762
PTRATIO     19.485238
PRICE       15.978571
dtype: float64
```

In [117]:
```
# Mean price of homes under $20,000 is $15,000. Homes below $20,000 comp
rise of 41 percent of the homes.
# It would be interesting to see the comparison of other attributes such
as distance to employment center, age,
# number of rooms etc.
```

Out[117]:

	CRIM	ZN	INDUS	RM	AGE	DIS	RAD	TAX	PTRATIO	PRICE
8	0.21124	12.5	7.87	5.631	100.0	6.0821	5	311	15.2	16.5
9	0.17004	12.5	7.87	6.004	85.9	6.5921	5	311	15.2	18.9
10	0.22489	12.5	7.87	6.377	94.3	6.3467	5	311	15.2	15.0
11	0.11747	12.5	7.87	6.009	82.9	6.2267	5	311	15.2	18.9
14	0.63796	0.0	8.14	6.096	84.5	4.4619	4	307	21.0	18.2

In [16]:
```
greaterthan20K = HousingData[HousingData['PRICE'] > 20]
greaterthan20K.mean()
```

Out[16]:
```
CRIM         0.937454
ZN          17.809278
INDUS        8.323608
RM           6.557938
AGE         55.718213
DIS          4.435600
RAD          5.649828
TAX        337.463918
PTRATIO     17.764487
PRICE       27.955670
dtype: float64
```

In [18]:
```
print("Difference in mean price is $"+str(np.mean(greaterthan20K['PRICE'
])-np.mean(lessthan20K['PRICE'])))
```

```
Difference in mean price is $12.877098674521356
```

In [20]:
```
differenceInDistance = np.mean(greaterthan20K['DIS'])-np.mean(lessthan20
K['DIS'])
if (differenceInDistance > 0):
    print("Difference to work center is higher for homes greater than $2
0K")
else:
```