1. (4pts) Give some examples for each of the four different types of attributes: Nominal; Ordinal; Interval; Ratio. For each of them, define a distance measure that is meaningful.

1. Nominal
    a. Barcode
    b. Color
    c. Distinctness
2. Ordinal
    a. Gold, Silver, Bronze Medals
    b. Company Employee Tier
    c. Distinctness & Order
3. Interval
    a. Time between events
    b. Age ranges
    c. Distinctness & Order & Meaningful Differences
4. Ratio
    a. Weight
    b. Poverty rate
    c. Distinctness & Order & Meaningful Differences & Meaningful Ratios

2. (2pts) Show that Extended Jaccard Coefficient (EJ) reduce to Jaccard coefficient if x and y are binary vectors.

$$EJ(\mathbf{x},\mathbf{y}) = \frac{\mathbf{x}\cdot\mathbf{y}}{\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - \mathbf{x}\cdot\mathbf{y}}$$

J = $(f_{11}) / (f_{01} + f_{10} + f_{11})$,

where f_{01} = the number of attributes where x was 0 and y was 1; f_{10} = the number of attributes where x was 1 and y was 0, f_{00} = the number of attributes where x was 0 and y was 0, f_{11} = the number of attributes where x was 1 and y was 1.

$$EJ(x,y) = \frac{x\cdot y}{||x||^2 + ||y||^2 - x\cdot y}$$

$$x \cdot y = f_{11}$$

This is because the dot product result is the sum of all the matches where x and y are 1 since otherwise they would be 0.

$$EJ(x,y) = \frac{f_{11}}{||x||^2+||y||^2-f_{11}}$$

$$||x||^2 + ||y||^2 = f_{01} + f_{11} + f_{10} + f_{11}$$

This is because the left side is the lengths of the two vectors which is equivalent to the number of unique elements in x and y and two times the duplicates since they are double counted.

$$EJ(x,y) = \frac{f_{11}}{f_{01}+f_{11}+f_{10}+f_{11}-f_{11}}$$

$$EJ(x,y) = \frac{f_{11}}{f_{01}+f_{11}+f_{10}}$$

And Thus they are equivalent.

3. (10pts) Mutual information of two variables X and Y can be defined as the difference between the summation of the entropy of X and Y, and their joint entropy. It can also be defined in three other ways as shown in the formula below:

$$\begin{aligned} \mathrm{I}(X;Y) &\equiv \mathrm{H}(X) - \mathrm{H}(X|Y) \\ &\equiv \mathrm{H}(Y) - \mathrm{H}(Y|X) \\ &\equiv \mathrm{H}(X) + \mathrm{H}(Y) - \mathrm{H}(X,Y) \\ &\equiv \mathrm{H}(X,Y) - \mathrm{H}(X|Y) - \mathrm{H}(Y|X) \end{aligned}$$

Prove all these four definitions are equivalent. You can assume X and Y are discrete variables.

We Know the following equation.

$$H(X,Y) = H(X|Y) + H(Y) = H(Y|X) + H(X)$$

2.

$$H(Y) - H(Y|X) = H(X) - H(Y|X)$$
$$H(X,Y) - H(X|Y) - H(Y|X)$$
$$H(X) - H(Y|X) = H(X) - H(Y|X)$$

3.

$$H(X) + H(Y) - H(X,Y) = H(X) - H(X|Y)$$
$$H(X) + H(Y) - H(Y) - H(X|Y)$$
$$H(X) - H(X|Y) = H(X) - H(X|Y)$$

4.

$$H(X,Y) - H(X|Y) - H(Y|X) = H(X) - H(X|Y)$$
$$H(Y|X) + H(X) - H(X|Y) - H(Y|X)$$
$$H(X) - H(X|Y) = H(X) - H(X|Y)$$

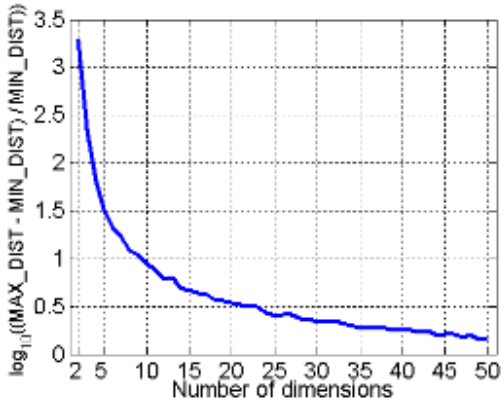
4. (10pts) Perform the following exercises to get you familiar with the steps in data analysis.

1) Read data from csv format (housing.csv) into DataFrame. Check the data information (e.g., structure, summary statistics) using functions such as head(), info(), describe().

2) Visualize the data (the nine numerical attributes) using hist(), boxplot(). 3) Draw the scatter plot of the first two variables (longitude and latitude). You can change the size each circle and its color based on the size of population and median house price (see references).

4) Draw the scatter plot of the two variables median house price and median income. Calculate the correlation coefficient of the two variables.

5. (4pts). Generate a random data matrix of the size 500X50. All of the values a uniformly distributed in the interval [0, 1]. For each index i from 2 to 50, use the first i variables to calculate the pairwise Euclid distance, we use Max_dist and Min_dist to represent the current max and min distance, respectively. Then calculate the value: log_10((Max_dist – Min_dist)/Min_dist) for each index i and draw the line diagram of these values (see slide 86 in Lecture 2, also shown here).



ForQ4&Q5, please export a PDF file from Jupyter Notebook that includes both code and results, and merge it with your answer to other questions and submit the final merged pdf file.

Notes Q4: if you need help with the functions, you can check the example code from the following references.

1. Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems By Géron, 2017. Chapter 2. End-to-End Machine Learning Project
2. http://www.cse.msu.edu/~ptan/dmbook/tutorials/tutorial4/tutorial4.html

HW #1

In [89]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

data = pd.read_csv('housing.csv')
```

Q4

1. Data Information

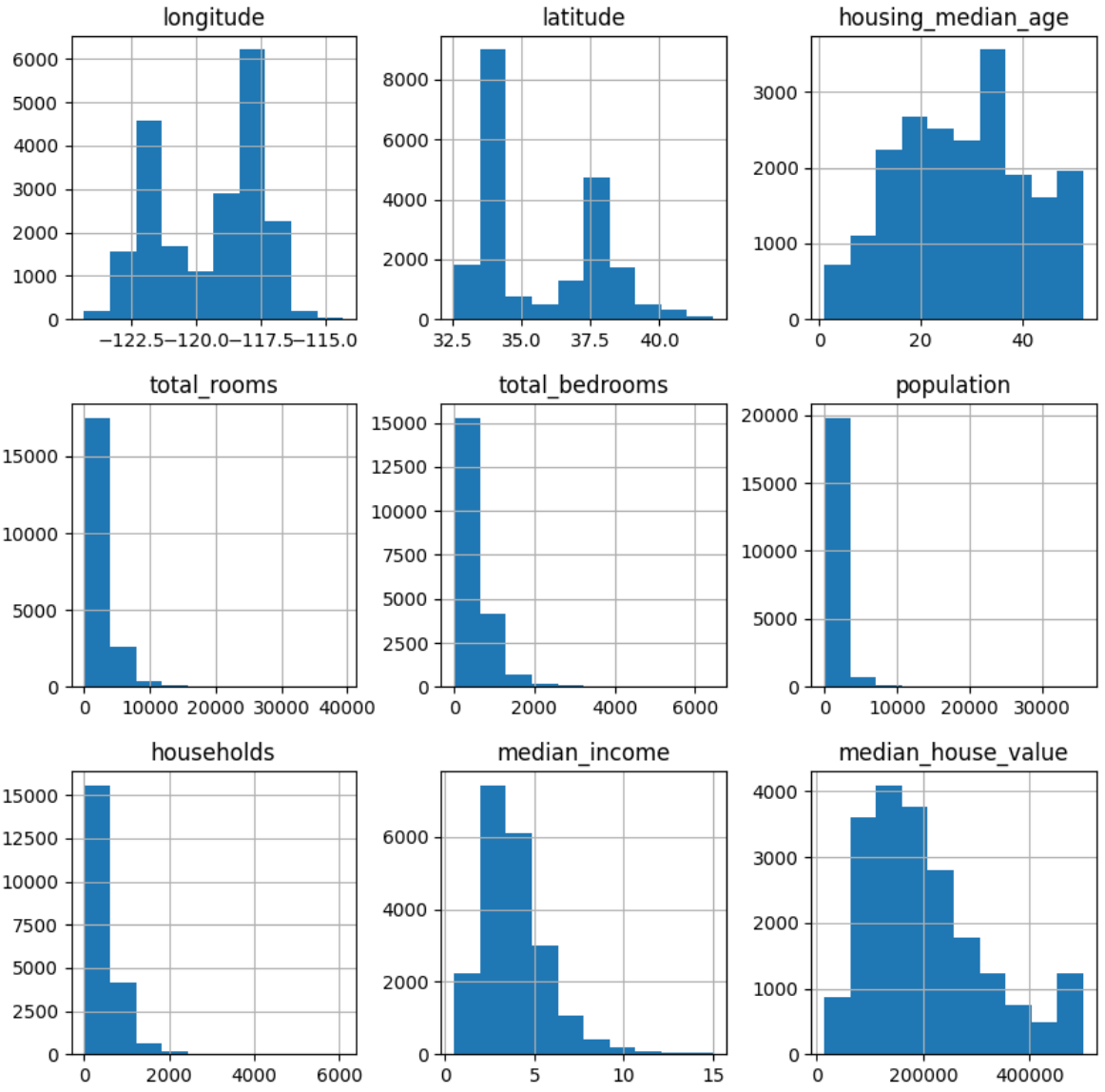
In [14]:
```
print('Data Head\n')
head = data.head()
print(head)
print('\nData Info\n')
info = data.info()
print('\nData Description\n')
describe = data.describe()
print(describe)
```

```
Data Head

   longitude  latitude  housing_median_age  total_rooms  total_bedrooms  \
0    -122.23     37.88                41.0        880.0           129.0
1    -122.22     37.86                21.0       7099.0          1106.0
2    -122.24     37.85                52.0       1467.0           190.0
3    -122.25     37.85                52.0       1274.0           235.0
4    -122.25     37.85                52.0       1627.0           280.0

   population  households  median_income  median_house_value ocean_proximi
ty
0       322.0       126.0         8.3252            452600.0        NEAR B
AY
1      2401.0      1138.0         8.3014            358500.0        NEAR B
AY
2       496.0       177.0         7.2574            352100.0        NEAR B
AY
3       558.0       219.0         5.6431            341300.0        NEAR B
AY
4       565.0       259.0         3.8462            342200.0        NEAR B
AY

Data Info

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20640 entries, 0 to 20639
Data columns (total 10 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   longitude           20640 non-null  float64
 1   latitude            20640 non-null  float64
 2   housing_median_age  20640 non-null  float64
 3   total_rooms         20640 non-null  float64
 4   total_bedrooms      20433 non-null  float64
 5   population          20640 non-null  float64
 6   households          20640 non-null  float64
 7   median_income       20640 non-null  float64
 8   median_house_value  20640 non-null  float64
 9   ocean_proximity     20640 non-null  object
dtypes: float64(9), object(1)
memory usage: 1.6+ MB

Data Description

          longitude      latitude  housing_median_age   total_rooms  \
count  20640.000000  20640.000000        20640.000000  20640.000000
mean    -119.569704     35.631861           28.639486   2635.763081
std        2.003532      2.135952           12.585558   2181.615252
min     -124.350000     32.540000            1.000000      2.000000
25%     -121.800000     33.930000           18.000000   1447.750000
50%     -118.490000     34.260000           29.000000   2127.000000
75%     -118.010000     37.710000           37.000000   3148.000000
max     -114.310000     41.950000           52.000000  39320.000000

       total_bedrooms    population    households  median_income  \
count    20433.000000  20640.000000  20640.000000   20640.000000
mean       537.870553   1425.476744    499.539680       3.870671
std        421.385070   1132.462122    382.329753       1.899822
min          1.000000      3.000000      1.000000       0.499900
25%        296.000000    787.000000    280.000000       2.563400
```

```
50%         435.000000   1166.000000    409.000000        3.534800
75%         647.000000   1725.000000    605.000000        4.743250
max        6445.000000  35682.000000   6082.000000       15.000100

       median_house_value
count        20640.000000
mean        206855.816909
std         115395.615874
min          14999.000000
25%         119600.000000
50%         179700.000000
75%         264725.000000
max         500001.000000
```

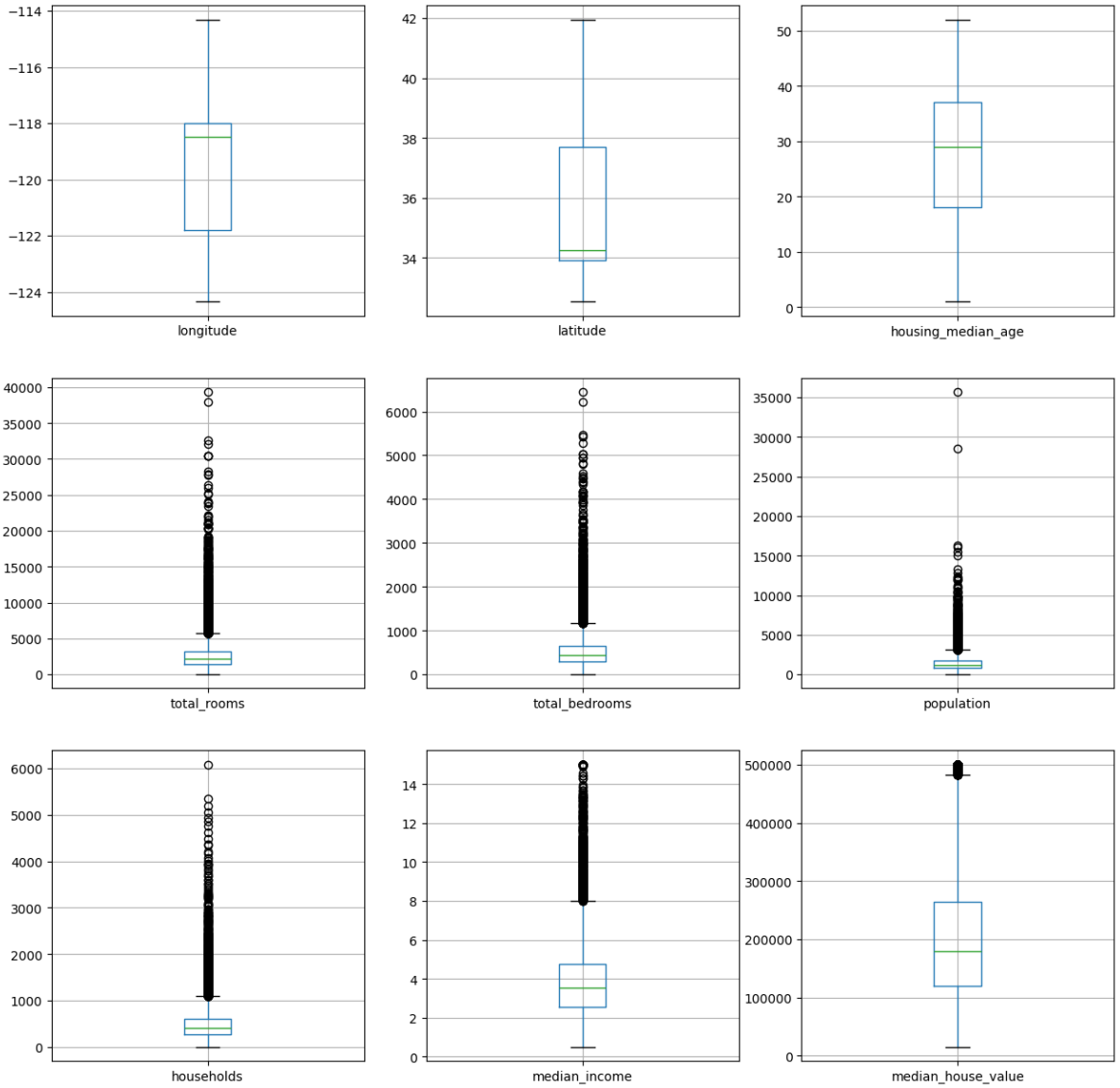
2. Visualize the Data

In [147…
```
data.hist(figsize=(10, 10))
plt.show()
```



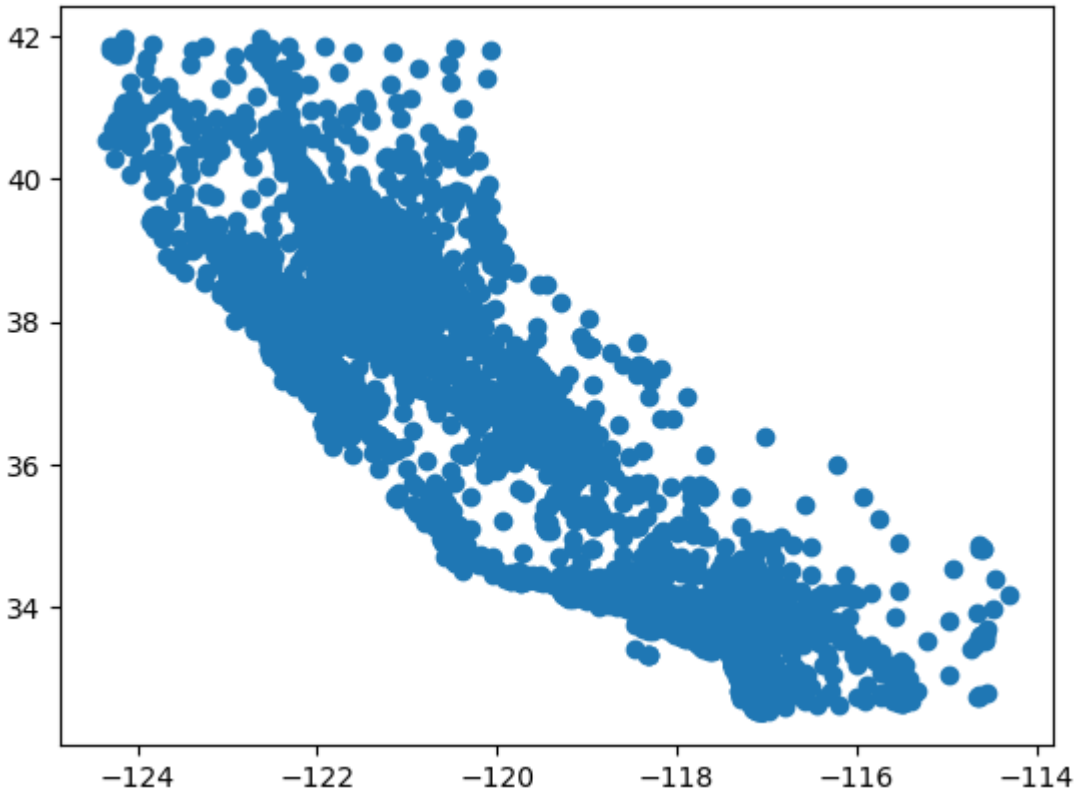
In [81]:
```
# Remove ocean_proxmitity
numerical_cols = data.columns.delete(9)
fig, axes = plt.subplots(3,3, figsize=(15, 15))
for index, col_name in enumerate(numerical_cols):
    row = index % 3
```

```
    col = int(index / 3)
    data.boxplot(col_name, ax=axes[col, row])
```

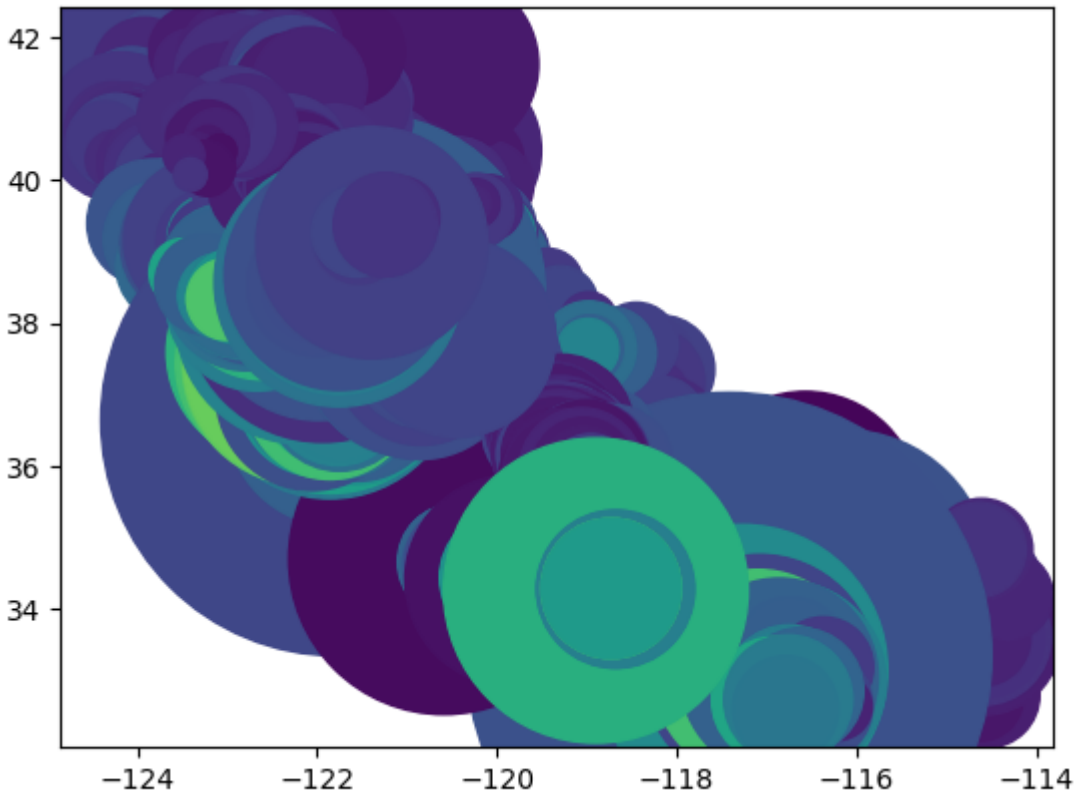


3. Longitude and Latitude Scatterplots

In [86]:
```
plt.scatter(data['longitude'], data['latitude'])
plt.show()
plt.scatter(data['longitude'], data['latitude'],data['population'], data[
```



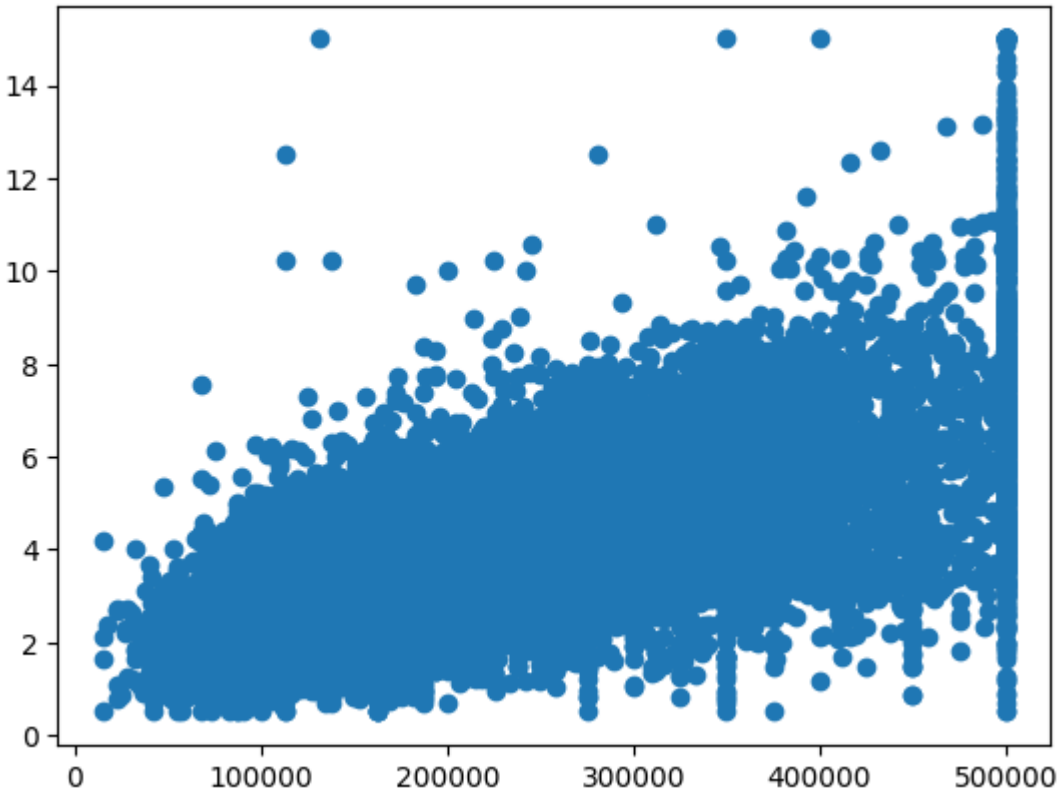
Out[86]: <matplotlib.collections.PathCollection at 0x78288d194f20>



4. Median House Price and Median Income Correlation

In [88]:
```
plt.scatter(data['median_house_value'],data['median_income'])

print(f'Correlation is {data['median_house_value'].corr(data['median_inco
```

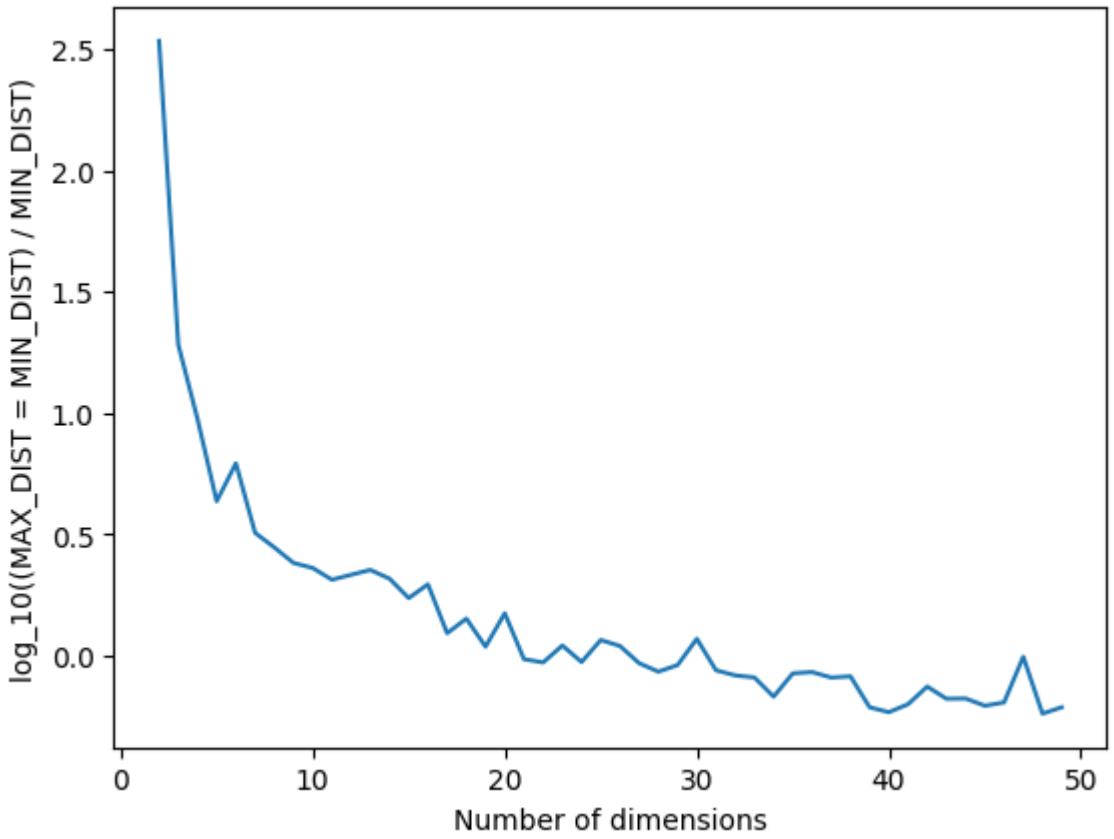
```
Correlation is 0.688075207958548
```



Q5

In [146…

```
import math
differences = []
MIN_DIM = 2
MAX_DIM = 50
SIZE = 500
DIM_RANGE = range(MIN_DIM, MAX_DIM)
for dim in DIM_RANGE:
    matrix_500_dim = np.random.rand(SIZE, dim)
    matrix_dim = np.random.rand(dim)
    diff = [np.linalg.norm(matrix_500_sliced - matrix_dim) for matrix_500

    max_diff = max(diff)
    min_diff = min(diff)
    log_diff = math.log10((max_diff-min_diff)/min_diff)
    differences.append(log_diff)
plt.plot(DIM_RANGE, differences)
plt.xlabel('Number of dimensions')
plt.ylabel('log_10((MAX_DIST = MIN_DIST) / MIN_DIST)')
plt.show()
```


log_10((MAX_DIST = MIN_DIST) / MIN_DIST)
2.5
2.0
1.5
1.0
0.5
0.0
0
10
20
30
40
50
Number of dimensions