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

In [2]:
```
df=pd.read_csv("10_USA_Housing.csv")
df
```

Out[2]:

	Avg. Area Income	Avg. Area House Age	Avg. Area Number of Rooms	Avg. Area Number of Bedrooms	Area Population	Price	Address
0	79545.458574	5.682861	7.009188	4.09	23086.800503	1.059034e+06	208 Michael Ferry Apt. 674\nLaurabury, NE 3701...
1	79248.642455	6.002900	6.730821	3.09	40173.072174	1.505891e+06	188 Johnson Views Suite 079\nLake Kathleen, CA...
2	61287.067179	5.865890	8.512727	5.13	36882.159400	1.058988e+06	9127 Elizabeth Stravenue\nDanieltown, WI 06482...
3	63345.240046	7.188236	5.586729	3.26	34310.242831	1.260617e+06	USS Barnett\nFPO AP 44820
4	59982.197226	5.040555	7.839388	4.23	26354.109472	6.309435e+05	USNS Raymond\nFPO AE 09386
...	...	...	...	...	...	...	...
4995	60567.944140	7.830362	6.137356	3.46	22837.361035	1.060194e+06	USNS Williams\nFPO AP 30153-7653
4996	78491.275435	6.999135	6.576763	4.02	25616.115489	1.482618e+06	PSC 9258, Box 8489\nAPO AA 42991-3352
4997	63390.686886	7.250591	4.805081	2.13	33266.145490	1.030730e+06	4215 Tracy Garden Suite 076\nJoshualand, VA 01...
4998	68001.331235	5.534388	7.130144	5.44	42625.620156	1.198657e+06	USS Wallace\nFPO AE 73316
4999	65510.581804	5.992305	6.792336	4.07	46501.283803	1.298950e+06	37778 George Ridges Apt. 509\nEast Holly, NV 2...

5000 rows × 7 columns

In [3]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5000 entries, 0 to 4999
Data columns (total 7 columns):
 #   Column                        Non-Null Count  Dtype  
---  ------                        --------------  -----  
 0   Avg. Area Income              5000 non-null   float64
 1   Avg. Area House Age           5000 non-null   float64
 2   Avg. Area Number of Rooms     5000 non-null   float64
 3   Avg. Area Number of Bedrooms  5000 non-null   float64
 4   Area Population               5000 non-null   float64
 5   Price                         5000 non-null   float64
 6   Address                       5000 non-null   object 
dtypes: float64(6), object(1)
memory usage: 273.6+ KB
```

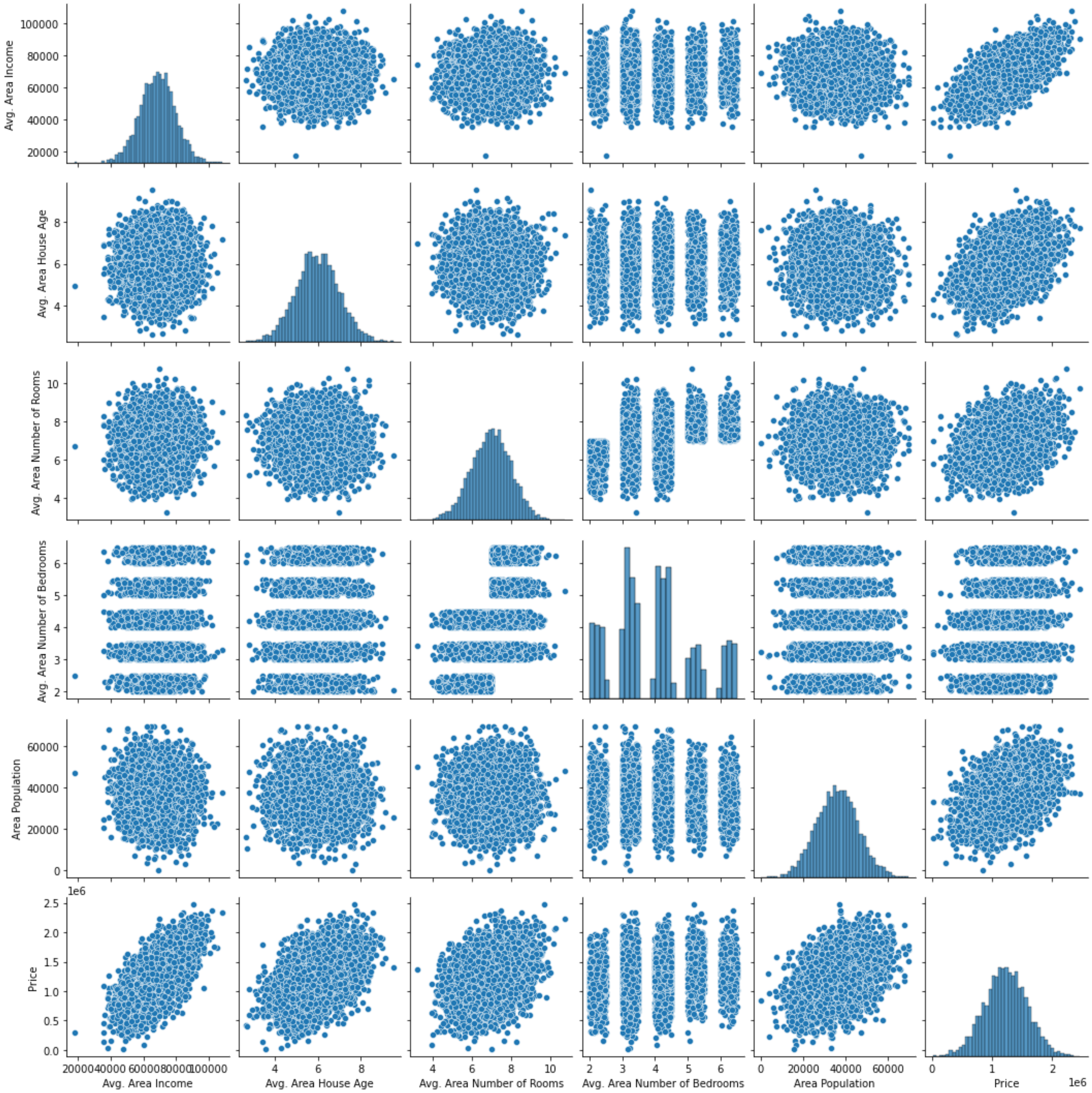
In [4]:
```
df.describe()
```

Out[4]:

	Avg. Area Income	Avg. Area House Age	Avg. Area Number of Rooms	Avg. Area Number of Bedrooms	Area Population	Price
count	5000.000000	5000.000000	5000.000000	5000.000000	5000.000000	5.000000e+03
mean	68583.108984	5.977222	6.987792	3.981330	36163.516039	1.232073e+06
std	10657.991214	0.991456	1.005833	1.234137	9925.650114	3.531176e+05
min	17796.631190	2.644304	3.236194	2.000000	172.610686	1.593866e+04
25%	61480.562388	5.322283	6.299250	3.140000	29403.928702	9.975771e+05
50%	68804.286404	5.970429	7.002902	4.050000	36199.406689	1.232669e+06
75%	75783.338666	6.650808	7.665871	4.490000	42861.290769	1.471210e+06
max	107701.748378	9.519088	10.759588	6.500000	69621.713378	2.469066e+06

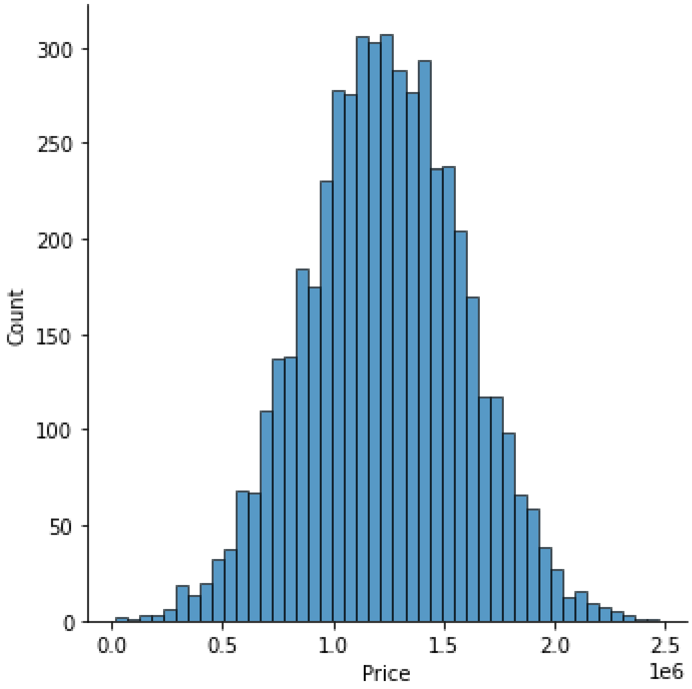
In [5]:
```
sns.pairplot(df)
```

Out[5]:
```
<seaborn.axisgrid.PairGrid at 0x1f2d5d1dd60>
```



```
In [6]: sns.displot(df['Price'])
```

```
Out[6]: <seaborn.axisgrid.FacetGrid at 0x1f2dbf8b760>
```



```
In [7]: df1=df.drop(['Address'],axis=1)
        df1
```

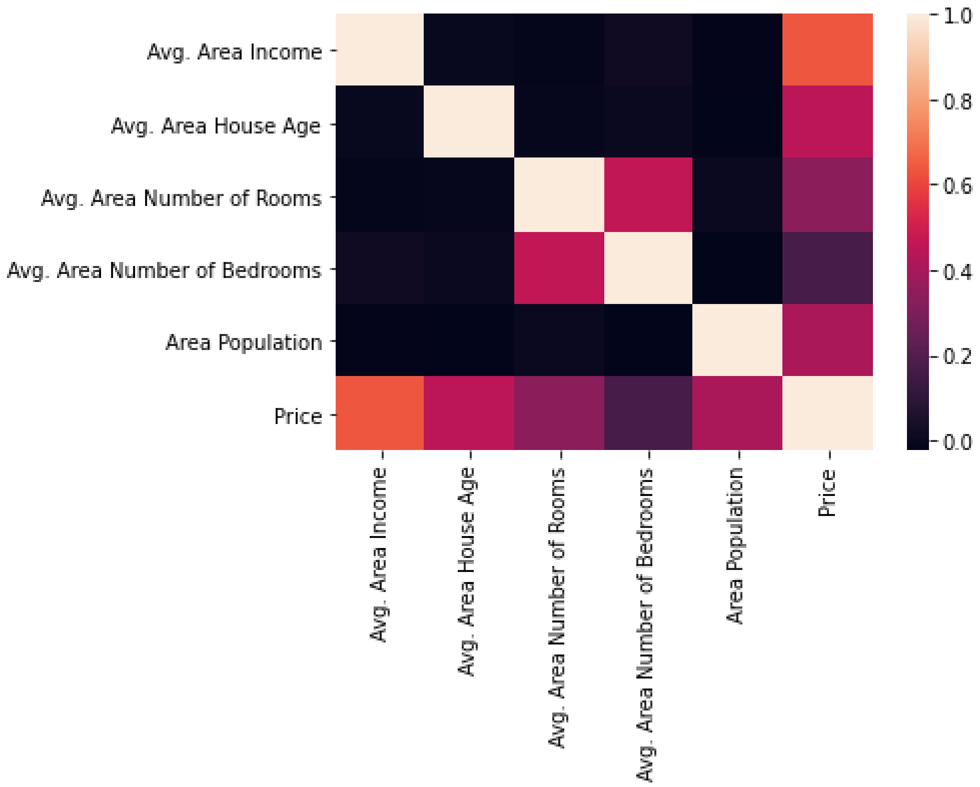
Out[7]:

	Avg. Area Income	Avg. Area House Age	Avg. Area Number of Rooms	Avg. Area Number of Bedrooms	Area Population	Price
0	79545.458574	5.682861	7.009188	4.09	23086.800503	1.059034e+06
1	79248.642455	6.002900	6.730821	3.09	40173.072174	1.505891e+06
2	61287.067179	5.865890	8.512727	5.13	36882.159400	1.058988e+06
3	63345.240046	7.188236	5.586729	3.26	34310.242831	1.260617e+06
4	59982.197226	5.040555	7.839388	4.23	26354.109472	6.309435e+05
...	...	...	...	...	...	...
4995	60567.944140	7.830362	6.137356	3.46	22837.361035	1.060194e+06
4996	78491.275435	6.999135	6.576763	4.02	25616.115489	1.482618e+06
4997	63390.686886	7.250591	4.805081	2.13	33266.145490	1.030730e+06
4998	68001.331235	5.534388	7.130144	5.44	42625.620156	1.198657e+06
4999	65510.581804	5.992305	6.792336	4.07	46501.283803	1.298950e+06

5000 rows × 6 columns

In [8]:
```
sns.heatmap(df1.corr())
```

Out[8]:
```
<AxesSubplot:>
```



In [9]:
```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
```

In [10]:
```
y=df['Price']
x=df1.drop(['Price'],axis=1)
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.3)
```

In [11]:
```
model=LinearRegression()
model.fit(x_train,y_train)
model.intercept_
```

Out[11]:
```
-2657812.6571027627
```

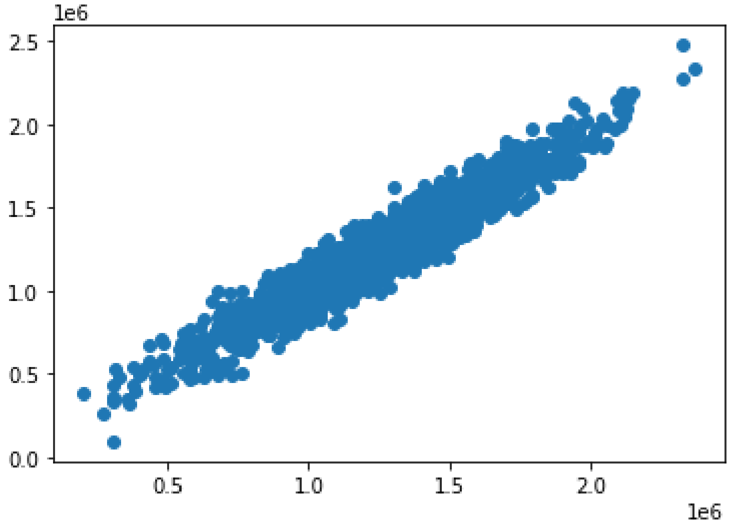
In [12]:
```
coeff=pd.DataFrame(model.coef_,x.columns,columns=["Coefficient"])
coeff
```

Out[12]:

	Coefficient
Avg. Area Income	21.540249
Avg. Area House Age	168198.049191
Avg. Area Number of Rooms	121591.585049
Avg. Area Number of Bedrooms	1573.129357
Area Population	15.229152

In [13]:
```
prediction=model.predict(x_test)
plt.scatter(y_test,prediction)
```

Out[13]:
```
<matplotlib.collections.PathCollection at 0x1f2de399070>
```



In [14]:
```
model.score(x_test,y_test)
```

Out[14]:
```
0.9169323598277197
```

In [56]:
```
df2=pd.read_csv("9_bottle.csv")
df2
```

```
C:\ProgramData\Anaconda3\lib\site-packages\IPython\core\interactiveshell.py:3165:
DtypeWarning: Columns (47,73) have mixed types.Specify dtype option on import or s
et low_memory=False.
  has_raised = await self.run_ast_nodes(code_ast.body, cell_name,
```

Out[56]:

	Cst_Cnt	Btl_Cnt	Sta_ID	Depth_ID	Depthm	T_degC	Salnty	O2ml_L	STheta	O2Sat	...
0	1	1	054.0 056.0	19-4903CR-HY-060-0930-05400560-0000A-3	0	10.500	33.4400	NaN	25.64900	NaN	...
1	1	2	054.0 056.0	19-4903CR-HY-060-0930-05400560-0008A-3	8	10.460	33.4400	NaN	25.65600	NaN	...
2	1	3	054.0 056.0	19-4903CR-HY-060-0930-05400560-0010A-7	10	10.460	33.4370	NaN	25.65400	NaN	...
3	1	4	054.0 056.0	19-4903CR-HY-060-0930-05400560-0019A-3	19	10.450	33.4200	NaN	25.64300	NaN	...
4	1	5	054.0 056.0	19-4903CR-HY-060-0930-05400560-0020A-7	20	10.450	33.4210	NaN	25.64300	NaN	...
...	...	...	...	...	...	...	...	...	...	...	...
864858	34404	864859	093.4 026.4	20-1611SR-MX-310-2239-09340264-0000A-7	0	18.744	33.4083	5.805	23.87055	108.74	...
864859	34404	864860	093.4 026.4	20-1611SR-MX-310-2239-09340264-0002A-3	2	18.744	33.4083	5.805	23.87072	108.74	...
864860	34404	864861	093.4 026.4	20-1611SR-MX-310-2239-09340264-0005A-3	5	18.692	33.4150	5.796	23.88911	108.46	...
864861	34404	864862	093.4 026.4	20-1611SR-MX-310-2239-09340264-0010A-3	10	18.161	33.4062	5.816	24.01426	107.74	...

	Cst_Cnt	Btl_Cnt	Sta_ID	Depth_ID	Depthm	T_degC	Salnty	O2ml_L	STheta	O2Sat	...
864862	34404	864863	093.4 026.4	20-1611SR-MX-310-2239-09340264-0015A-3	15	17.533	33.3880	5.774	24.15297	105.66	...

864863 rows × 74 columns

In [57]:
```
df2.isna().sum()
```

Out[57]:
```
Cst_Cnt                      0
Btl_Cnt                      0
Sta_ID                       0
Depth_ID                     0
Depthm                       0
                         ...  
TA1                     862779
TA2                     864629
pH2                     864853
pH1                     864779
DIC Quality Comment     864808
Length: 74, dtype: int64
```

In [58]:
```
df2.columns
```

Out[58]:
```
Index(['Cst_Cnt', 'Btl_Cnt', 'Sta_ID', 'Depth_ID', 'Depthm', 'T_degC',
       'Salnty', 'O2ml_L', 'STheta', 'O2Sat', 'Oxy_µmol/Kg', 'BtlNum',
       'RecInd', 'T_prec', 'T_qual', 'S_prec', 'S_qual', 'P_qual', 'O_qual',
       'SThtaq', 'O2Satq', 'ChlorA', 'Chlqua', 'Phaeop', 'Phaqua', 'PO4uM',
       'PO4q', 'SiO3uM', 'SiO3qu', 'NO2uM', 'NO2q', 'NO3uM', 'NO3q', 'NH3uM',
       'NH3q', 'C14As1', 'C14A1p', 'C14A1q', 'C14As2', 'C14A2p', 'C14A2q',
       'DarkAs', 'DarkAp', 'DarkAq', 'MeanAs', 'MeanAp', 'MeanAq', 'IncTim',
       'LightP', 'R_Depth', 'R_TEMP', 'R_POTEMP', 'R_SALINITY', 'R_SIGMA',
       'R_SVA', 'R_DYNHT', 'R_O2', 'R_O2Sat', 'R_SIO3', 'R_PO4', 'R_NO3',
       'R_NO2', 'R_NH4', 'R_CHLA', 'R_PHAEO', 'R_PRES', 'R_SAMP', 'DIC1',
       'DIC2', 'TA1', 'TA2', 'pH2', 'pH1', 'DIC Quality Comment'],
      dtype='object')
```

In [59]:
```
df3=df2.drop(['DIC2', 'TA1', 'TA2', 'pH2', 'pH1', 'DIC Quality Comment'],axis=1)
```

In [ ]:
```

```

In [ ]:
```

```

In [60]:

```
df4=df3.drop(['BtlNum', 'T_qual', 'S_qual', 'O_qual', 'SThtaq', 'NH3uM', 'C14As1',
       'C14A1p', 'C14As2', 'C14A2p', 'DarkAs', 'DarkAp', 'MeanAs', 'MeanAp',
       'IncTim', 'LightP', 'R_NH4', 'R_SAMP', 'DIC1'],axis=1)
```

In [61]:
```
df4.isna().sum()/len(df4)*100
```

Out[61]:
```
Cst_Cnt         0.000000
Btl_Cnt         0.000000
Sta_ID          0.000000
Depth_ID        0.000000
Depthm          0.000000
T_degC          1.267600
Salnty          5.475318
O2ml_L         19.501586
STheta          6.092179
O2Sat          23.540029
Oxy_µmol/Kg    23.540723
RecInd          0.000000
T_prec          1.267600
S_prec          5.475318
P_qual         22.096910
O2Satq         74.817168
ChlorA         73.952869
Chlqua         26.096272
Phaeop         73.952984
Phaqua         26.095809
PO4uM          52.210119
PO4q           47.762131
SiO3uM         59.058140
SiO3qu         40.930991
NO2uM          60.967691
NO2q           38.779437
NO3uM          60.987694
NO3q           38.726365
NH3q            6.540227
C14A1q          1.879835
C14A2q          1.877754
DarkAq          2.823915
MeanAq          2.824031
R_Depth         0.000000
R_TEMP          1.267600
R_POTEMP        5.324196
R_SALINITY      5.475318
R_SIGMA         6.111488
R_SVA           6.101660
R_DYNHT         5.394727
R_O2           19.501586
R_O2Sat        22.941784
R_SIO3         59.057215
R_PO4          52.209194
R_NO3          60.986769
R_NO2          60.966766
R_CHLA         73.952406
R_PHAEO        73.952522
R_PRES          0.000000
dtype: float64
```

In [ ]:

```

```

In [62]:

```
df4.describe()
```

Out[62]:

	Cst_Cnt	Btl_Cnt	Depthm	T_degC	Salnty	O2ml_L	
count	864863.000000	864863.000000	864863.000000	853900.000000	817509.000000	696201.000000	812
mean	17138.790958	432432.000000	226.831951	10.799677	33.840350	3.392468	
std	10240.949817	249664.587267	316.050259	4.243825	0.461843	2.073256	
min	1.000000	1.000000	0.000000	1.440000	28.431000	-0.010000	
25%	8269.000000	216216.500000	46.000000	7.680000	33.488000	1.360000	
50%	16848.000000	432432.000000	125.000000	10.060000	33.863000	3.440000	
75%	26557.000000	648647.500000	300.000000	13.880000	34.196900	5.500000	
max	34404.000000	864863.000000	5351.000000	31.140000	37.034000	11.130000	

8 rows × 47 columns

In [63]:

```
df4.shape
```

Out[63]:

```
(864863, 49)
```

In [64]:

```
df5=df4.iloc[0:5000,0:]
df5.shape
```

Out[64]:

```
(5000, 49)
```

In [65]:

```
per=df5.isna().sum()/len(df5)*100
per1=pd.DataFrame(per,df5.columns,)
per1
```

Out[65]:

	0
Cst_Cnt	0.00
Btl_Cnt	0.00
Sta_ID	0.00
Depth_ID	0.00
Depthm	0.00
T_degC	0.40
Salnty	3.04
O2ml_L	43.80
STheta	3.34
O2Sat	45.74
Oxy_µmol/Kg	45.74
RecInd	0.00
T_prec	0.40
S_prec	3.04
P_qual	0.00
O2Satq	52.88
ChlorA	100.00
Chlqua	0.00
Phaeop	100.00
Phaqua	0.00
PO4uM	79.08
PO4q	20.92
SiO3uM	100.00
SiO3qu	0.00
NO2uM	100.00
NO2q	0.00
NO3uM	100.00
NO3q	0.00
NH3q	0.00
C14A1q	0.00
C14A2q	0.00
DarkAq	0.00
MeanAq	0.00
R_Depth	0.00
R_TEMP	0.40
R_POTEMP	4.50

	0
R_SALINITY	3.04
R_SIGMA	5.62
R_SVA	5.62
R_DYNHT	4.28
R_O2	43.80
R_O2Sat	46.20
R_SIO3	100.00
R_PO4	79.08
R_NO3	100.00
R_NO2	100.00
R_CHLA	100.00
R_PHAEO	100.00
R_PRES	0.00

```
In [66]: pr=per1[per1[0]>75].index
         pr
```

```
Out[66]: Index(['ChlorA', 'Phaeop', 'PO4uM', 'SiO3uM', 'NO2uM', 'NO3uM', 'R_SIO3',
                'R_PO4', 'R_NO3', 'R_NO2', 'R_CHLA', 'R_PHAEO'],
               dtype='object')
```

In [90]:

```
df6=df5.drop(['ChlorA', 'Phaeop', 'PO4uM', 'SiO3uM', 'NO2uM', 'NO3uM', 'R_SIO3',
       'R_PO4', 'R_NO3', 'R_NO2', 'R_CHLA', 'R_PHAEO'],axis=1)
df6.isna().sum()/len(df5)*100
```

Out[90]:

```
Cst_Cnt          0.00
Btl_Cnt          0.00
Sta_ID           0.00
Depth_ID         0.00
Depthm           0.00
T_degC           0.40
Salnty           3.04
O2ml_L          43.80
STheta           3.34
O2Sat           45.74
Oxy_µmol/Kg     45.74
RecInd           0.00
T_prec           0.40
S_prec           3.04
P_qual           0.00
O2Satq          52.88
Chlqua           0.00
Phaqua           0.00
PO4q            20.92
SiO3qu           0.00
NO2q             0.00
NO3q             0.00
NH3q             0.00
C14A1q           0.00
C14A2q           0.00
DarkAq           0.00
MeanAq           0.00
R_Depth          0.00
R_TEMP           0.40
R_POTEMP         4.50
R_SALINITY       3.04
R_SIGMA          5.62
R_SVA            5.62
R_DYNHT          4.28
R_O2            43.80
R_O2Sat         46.20
R_PRES           0.00
dtype: float64
```

In [91]:

```

nul_col = df6.columns[df6.isnull().any()].tolist()
nul_col
```

Out[91]:

```
['T_degC',
 'Salnty',
 'O2ml_L',
 'STheta',
 'O2Sat',
 'Oxy_µmol/Kg',
 'T_prec',
 'S_prec',
 'O2Satq',
 'PO4q',
 'R_TEMP',
 'R_POTEMP',
 'R_SALINITY',
 'R_SIGMA',
 'R_SVA',
 'R_DYNHT',
 'R_O2',
 'R_O2Sat']
```

In [88]:

```
#for i in nul_col:
    #df6[i]= df6.fillna(df6[i].mean())
```

In [89]:

```
#df6
```

Out[89]:

	Cst_Cnt	Btl_Cnt	Sta_ID	Depth_ID	Depthm	T_degC	Salnty	O2ml_L	STheta	O2Sat	...	R_De
0	1	1	054.0 056.0	19-4903CR-HY-060-0930-05400560-0000A-3	0	1	1	1	1	1	...	
1	1	2	054.0 056.0	19-4903CR-HY-060-0930-05400560-0008A-3	8	1	1	1	1	1	...	
2	1	3	054.0 056.0	19-4903CR-HY-060-0930-05400560-0010A-7	10	1	1	1	1	1	...	1
3	1	4	054.0 056.0	19-4903CR-HY-060-0930-05400560-0019A-3	19	1	1	1	1	1	...	1
4	1	5	054.0 056.0	19-4903CR-HY-060-0930-05400560-0020A-7	20	1	1	1	1	1	...	2
...	...	...	...	...	...	...	...	...	...	...	...	
4995	165	4996	092.0 098.0	19-4904NS-HY-102-1342-09200980-0099A-3	99	165	165	165	165	165	...	9
4996	165	4997	092.0 098.0	19-4904NS-HY-102-1342-09200980-0100A-7	100	165	165	165	165	165	...	10
4997	165	4998	092.0 098.0	19-4904NS-HY-102-1342-09200980-0125A-7	125	165	165	165	165	165	...	12
4998	165	4999	092.0 098.0	19-4904NS-HY-102-1342-09200980-0149A-3	149	165	165	165	165	165	...	14

	Cst_Cnt	Btl_Cnt	Sta_ID	Depth_ID	Depthm	T_degC	Salnty	O2ml_L	STheta	O2Sat	...	R_De
4999	165	5000	092.0 098.0	19-4904NS-HY-102-1342-09200980-0150A-7	150	165	165	165	165	165	...	15

5000 rows × 37 columns

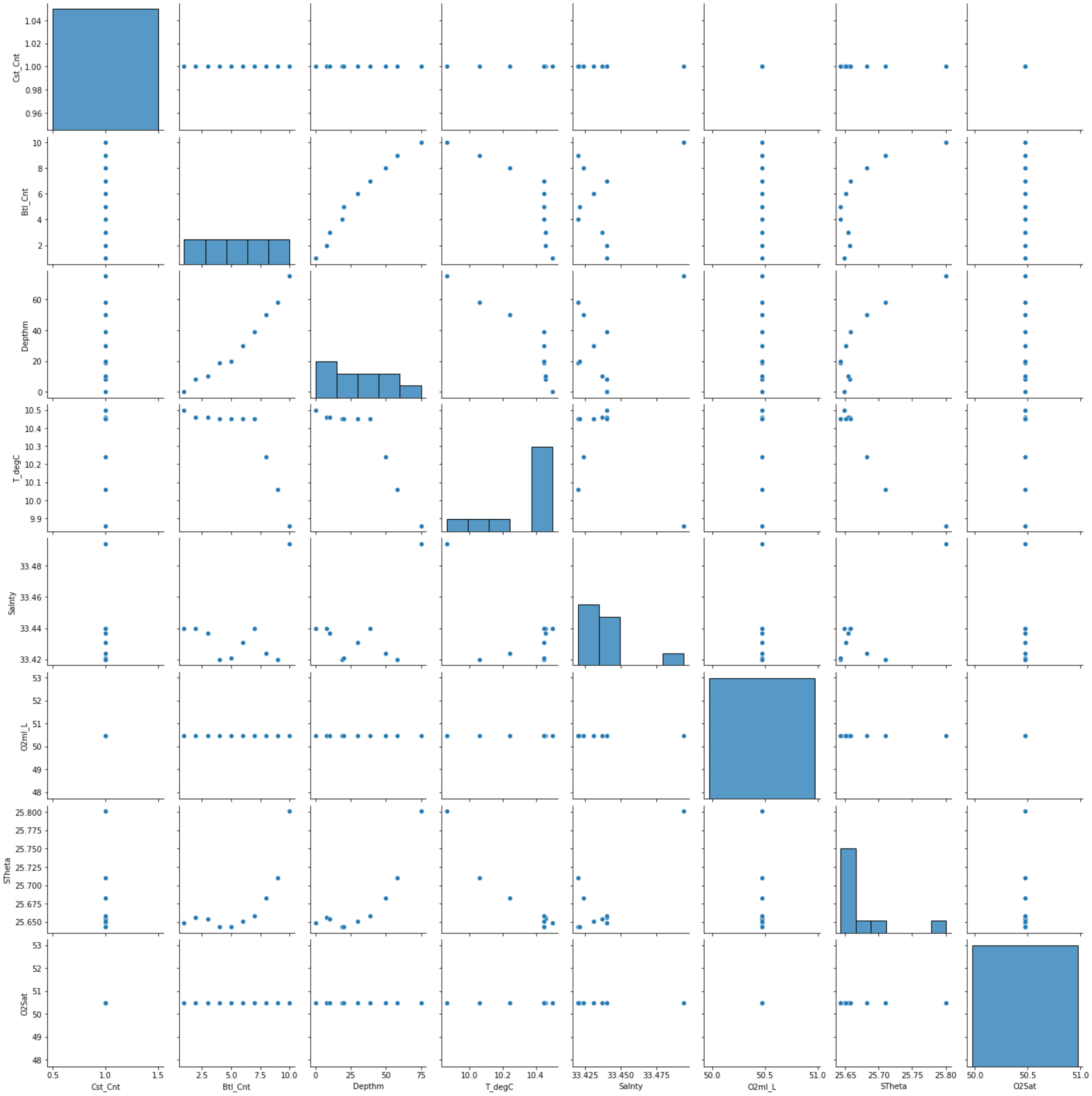
In [94]:
```
df6.fillna(df6["T_degC"].mean())
df6.fillna(df6["Salnty"].mean())
df6.fillna(df6["O2ml_L"].mean())
df6.fillna(df6["STheta"].mean())
df6.fillna(df6["O2Sat"].median())
df6.fillna(df6["Oxy_µmol/Kg"].mean())
df6.fillna(df6["T_prec"].mean())
df6.fillna(df6["S_prec"].mean())
df6.fillna(df6["O2Satq"].mean())
df6.fillna(df6["PO4q"].mean())
df6.fillna(df6["R_TEMP"].mean())
df6.fillna(df6["R_POTEMP"].mean())
df6.fillna(df6["R_SALINITY"].mean())
df6.fillna(df6["R_SIGMA"].mean())
df6.fillna(df6["R_SVA"].mean())
df6.fillna(df6["R_DYNHT"].mean())
df6.fillna(df6["R_O2"].mean())
df6=df6.fillna(df6["R_O2Sat"].mean())
```

In [95]:
```
df6.isna().sum()
```

Out[95]:
```
Cst_Cnt        0
Btl_Cnt        0
Sta_ID         0
Depth_ID       0
Depthm         0
T_degC         0
Salnty         0
O2ml_L         0
STheta         0
O2Sat          0
Oxy_µmol/Kg    0
RecInd         0
T_prec         0
S_prec         0
P_qual         0
O2Satq         0
Chlqua         0
Phaqua         0
PO4q           0
SiO3qu         0
NO2q           0
NO3q           0
NH3q           0
C14A1q         0
C14A2q         0
DarkAq         0
MeanAq         0
R_Depth        0
R_TEMP         0
R_POTEMP       0
R_SALINITY     0
R_SIGMA        0
R_SVA          0
R_DYNHT        0
R_O2           0
R_O2Sat        0
R_PRES         0
dtype: int64
```

In [98]:
```
df7=df6.iloc[:10,:10]
sns.pairplot(df7)
```

Out[98]:
```
<seaborn.axisgrid.PairGrid at 0x1f287b2eac0>
```

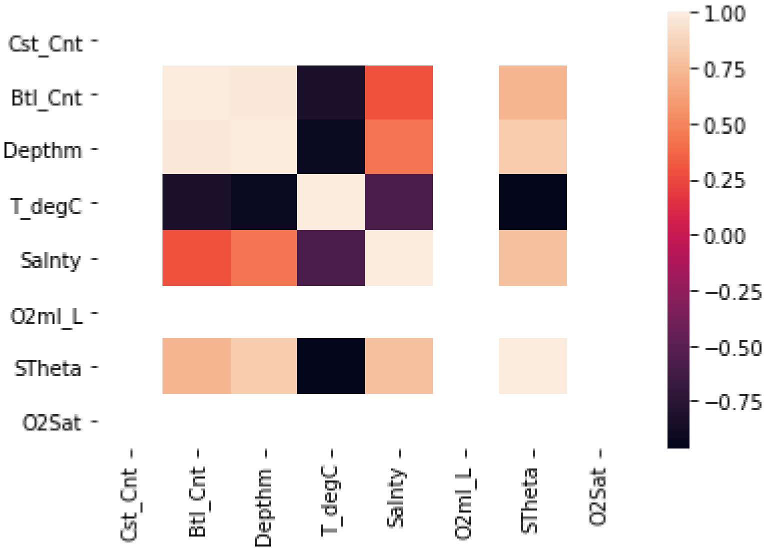


In [97]:

```
sns.heatmap(df7.corr())
```

Out[97]:

```
<AxesSubplot:>
```



In [107]:

```
df6=df6.drop(["Sta_ID","Depth_ID"],axis=1)
```

In [109]:

```
x=df6.drop(["R_O2"],axis=1)
y=df6["R_O2"]
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.3)
model1=LinearRegression()
model1.fit(x_train,y_train)
model1.intercept_
```

Out[109]:

```
-4.618527782440651e-14
```

In [110]:

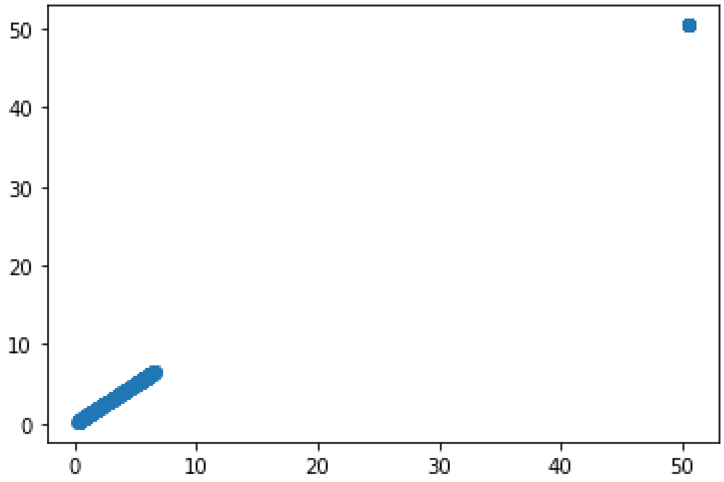
```
coeff=pd.DataFrame(model1.coef_,x.columns,columns=["Coefficient"])
coeff
```

Out[110]:

	Coefficient
Cst_Cnt	-1.900857e-14
Btl_Cnt	5.568462e-16
Depthm	2.371912e-15
T_degC	7.851962e-16
Salnty	1.374703e-14
O2ml_L	1.000000e+00
STheta	-9.098279e-15
O2Sat	2.150917e-15
Oxy_µmol/Kg	7.801648e-16
RecInd	3.879562e-16
T_prec	3.818522e-15
S_prec	-3.387600e-15
P_qual	-1.942890e-16
O2Satq	-3.325655e-16
Chlqua	2.775558e-17
Phaqua	1.387779e-17
PO4q	1.632673e-16
SiO3qu	0.000000e+00
NO2q	-1.734723e-18
NO3q	0.000000e+00
NH3q	0.000000e+00
C14A1q	0.000000e+00
C14A2q	0.000000e+00
DarkAq	0.000000e+00
MeanAq	0.000000e+00
R_Depth	2.433526e-15
R_TEMP	1.450847e-16
R_POTEMP	9.528765e-16
R_SALINITY	1.280675e-14
R_SIGMA	-1.068890e-15
R_SVA	-2.690356e-15
R_DYNHT	1.852546e-16
R_O2Sat	1.030526e-16
R_PRES	-5.059402e-15

In [113]:
```
pred=model1.predict(x_test)
plt.scatter(y_test,pred)
```

Out[113]:
```
<matplotlib.collections.PathCollection at 0x1f28a3286d0>
```



In [116]:
```
model1.score(x_test,y_test)
```

Out[116]:
```
1.0
```