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

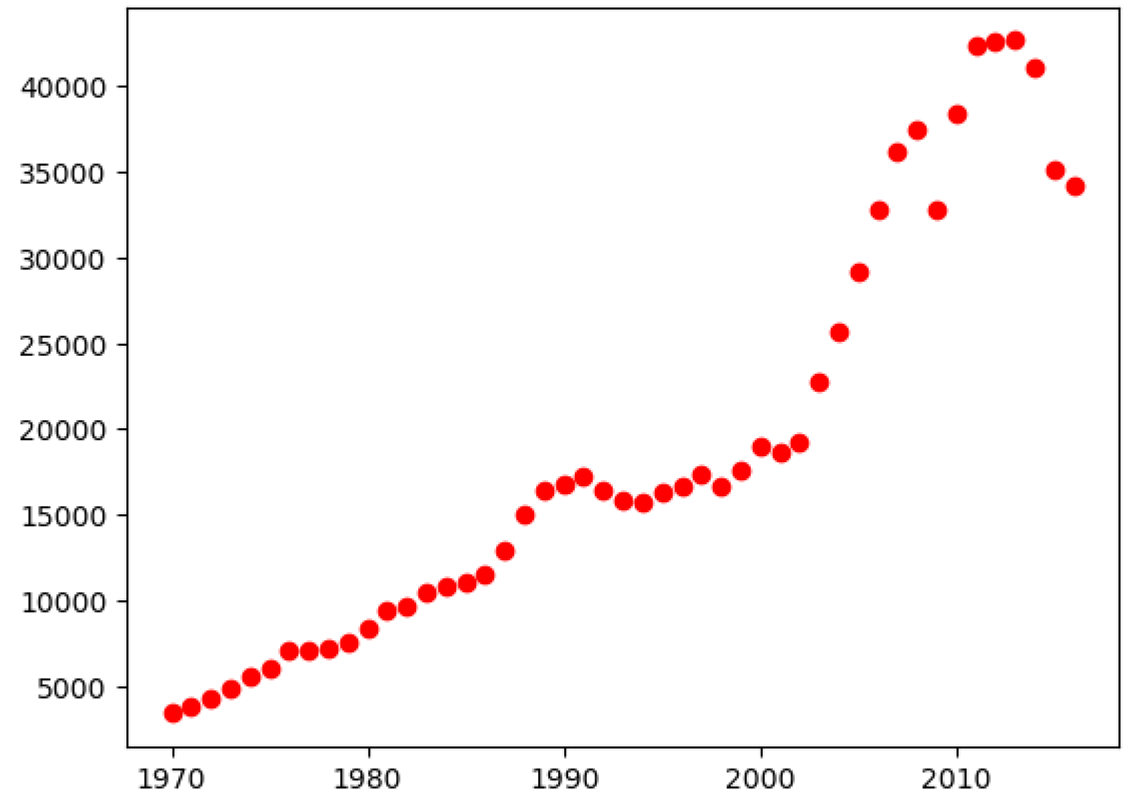
In [2]:
```python
dataset = pd.read_csv("canada_per_capita_income.csv")
```

In [3]:
```python
x = dataset.iloc[:,0:1].values
x
```

Out[3]:
```
array([[1970],
       [1971],
       [1972],
       [1973],
       [1974],
       [1975],
       [1976],
       [1977],
       [1978],
       [1979],
       [1980],
       [1981],
       [1982],
       [1983],
       [1984],
       [1985],
       [1986],
       [1987],
       [1988],
       [1989],
       [1990],
       [1991],
       [1992],
       [1993],
       [1994],
       [1995],
       [1996],
       [1997],
       [1998],
       [1999],
       [2000],
       [2001],
       [2002],
       [2003],
       [2004],
       [2005],
       [2006],
       [2007],
       [2008],
       [2009],
       [2010],
       [2011],
       [2012],
       [2013],
       [2014],
       [2015],
       [2016]])
```

In [4]:
```python
y = dataset.iloc[:, -1].values
y
```

Out[4]:
```
array([ 3399.299037,  3768.297935,  4251.175484,  4804.463248,
        5576.514583,  5998.144346,  7062.131392,  7100.12617 ,
        7247.967035,  7602.912681,  8355.96812 ,  9434.390652,
        9619.438377, 10416.53659 , 10790.32872 , 11018.95585 ,
       11482.89153 , 12974.80662 , 15080.28345 , 16426.72548 ,
       16838.6732  , 17266.09769 , 16412.08309 , 15875.58673 ,
       15755.82027 , 16369.31725 , 16699.82668 , 17310.75775 ,
       16622.67187 , 17581.02414 , 18987.38241 , 18601.39724 ,
       19232.17556 , 22739.42628 , 25719.14715 , 29198.05569 ,
       32738.2629  , 36144.48122 , 37446.48609 , 32755.17682 ,
       38420.52289 , 42334.71121 , 42665.25597 , 42676.46837 ,
       41039.8936  , 35175.18898 , 34229.19363 ])
```

In [5]:
```
# check the distribution of data is linear or not
plt.scatter(x,y,color='red')
```
Out[5]:
```
<matplotlib.collections.PathCollection at 0x7fd2694c5250>
```


In [6]:
```
#take 70% of data for training and 30% for testing
#random_state indicates the random seed used in splitting the data

from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(x,y, test_size=0.3,
```
In [7]:
```
#train using SLR
from sklearn.linear_model import LinearRegression
reg = LinearRegression()
reg.fit(x_train, y_train)
```
Out[7]:
```
▾ LinearRegression
LinearRegression()
```
In [8]:
```
#make prediction based on the test data
y_pred = reg.predict(x_test)
y_pred
```
Out[8]:
```
array([22865.78838995, 21142.72401552,  3912.08027117, 13388.93433056,
       18558.12745387, 15111.998705  ,   465.95152231, 34065.70682378,
        1327.48370952, 23727.32057717,  2189.01589674, 30619.57807491,
       -1257.11285213, 16835.06307943,  3050.54808396])
```
In [9]:
```
#find the r-squared value by comparing test data and predicted data
from sklearn.metrics import r2_score
r2_score(y_test, y_pred)
```
Out[9]:
```
0.8433026110551844
```
In [10]:
```
#another way to know the score of a linear reg model
reg.score(x_test, y_test)
```
Out[10]:
```
0.8433026110551844
```
In [11]:
```
reg.predict([[2022],[2023]])
```
Out[11]:
```
array([43542.56088317, 44404.09307038])
```

In [14]:
```
plt.scatter(x_train, y_train, color='red')
plt.plot(x_train, reg.predict(x_train),color='blue')
plt.title("Per Capita Income of Canada")
plt.xlabel('Year')
plt.ylabel('Per Capita Income')
plt.show()
```