In [1]:
```
import pandas as pd
```

In [2]:
```
url = "SquareFeet_Data2.csv"
data = pd.read_csv(url)
data
```

Out[2]:

	SquareFeet	Density	City	price
0	850	low	CityA	467500
1	779	medium	CityB	363014
2	990	medium	CityA	594000
3	665	medium	CityC	266000
4	550	medium	CityC	220000
5	880	medium	CityB	478720
6	567	low	CityB	264222
7	1020	low	CityB	497760
8	2067	high	CityC	756522
9	577	high	CityA	375050
10	989	medium	CityA	581532
11	720	low	CityC	165600
12	585	high	CityA	321750
13	656	medium	CityC	196800
14	788	high	CityC	253736
15	1222	high	CityB	596336
16	565	high	CityA	326005
17	844	high	CityC	196652
18	744	low	CityA	415152
19	1356	high	CityB	737664
20	1555	low	CityB	622000
21	2000	high	CityC	560000
22	647	low	CityA	355850
23	769	low	CityC	196095
24	855	medium	CityB	470250
25	900	medium	CityA	509400
26	456	medium	CityC	151848
27	669	low	CityB	347880
28	1899	high	CityA	1044450
29	633	low	CityC	212055

	SquareFeet	Density	City	price
30	890	medium	CityB	400500
31	946	low	CityC	272448
32	1235	low	CityA	679250

In [3]:
```
#pre-data processing
#convert sqft to m^2
#convert price to k

data["SquareFeet"] = data["SquareFeet"] * 0.092903
data['price'] = data['price'] / 1000
```

In [4]:
```
data["Density"] = data["Density"].map({"low": 0, "medium": 1, "high" : 2})
data
```

Out[4]:

	SquareFeet	Density	City	price
0	78.967550	0	CityA	467.500
1	72.371437	1	CityB	363.014
2	91.973970	1	CityA	594.000
3	61.780495	1	CityC	266.000
4	51.096650	1	CityC	220.000
5	81.754640	1	CityB	478.720
6	52.676001	0	CityB	264.222
7	94.761060	0	CityB	497.760
8	192.030501	2	CityC	756.522
9	53.605031	2	CityA	375.050
10	91.881067	1	CityA	581.532
11	66.890160	0	CityC	165.600
12	54.348255	2	CityA	321.750
13	60.944368	1	CityC	196.800
14	73.207564	2	CityC	253.736
15	113.527466	2	CityB	596.336
16	52.490195	2	CityA	326.005
17	78.410132	2	CityC	196.652
18	69.119832	0	CityA	415.152
19	125.976468	2	CityB	737.664
20	144.464165	0	CityB	622.000
21	185.806000	2	CityC	560.000
22	60.108241	0	CityA	355.850
23	71.442407	0	CityC	196.095
24	79.432065	1	CityB	470.250
25	83.612700	1	CityA	509.400
26	42.363768	1	CityC	151.848
27	62.152107	0	CityB	347.880
28	176.422797	2	CityA	1044.450
29	58.807599	0	CityC	212.055

	SquareFeet	Density	City	price
30	82.683670	1	CityB	400.500
31	87.886238	0	CityC	272.448
32	114.735205	0	CityA	679.250

One Hot Encoding

In [5]:
```
#Convert City to CityA, CityB, CityC
#not need CityC is because if CityA and CityB are 0. it means it is CityC
```

In [6]:
```
from sklearn.preprocessing import OneHotEncoder

onehot_encoder = OneHotEncoder()
onehot_encoder.fit(data[["City"]])
city_encoded = onehot_encoder.transform(data[["City"]]).toarray()

data[["CityA","CityB","CityC"]] = city_encoded
data = data.drop(["City","CityC"],axis=1)
data
```

Out[6]:

	SquareFeet	Density	price	CityA	CityB
0	78.967550	0	467.500	1.0	0.0
1	72.371437	1	363.014	0.0	1.0
2	91.973970	1	594.000	1.0	0.0
3	61.780495	1	266.000	0.0	0.0
4	51.096650	1	220.000	0.0	0.0
5	81.754640	1	478.720	0.0	1.0
6	52.676001	0	264.222	0.0	1.0
7	94.761060	0	497.760	0.0	1.0
8	192.030501	2	756.522	0.0	0.0
9	53.605031	2	375.050	1.0	0.0
10	91.881067	1	581.532	1.0	0.0
11	66.890160	0	165.600	0.0	0.0
12	54.348255	2	321.750	1.0	0.0
13	60.944368	1	196.800	0.0	0.0
14	73.207564	2	253.736	0.0	0.0
15	113.527466	2	596.336	0.0	1.0
16	52.490195	2	326.005	1.0	0.0
17	78.410132	2	196.652	0.0	0.0
18	69.119832	0	415.152	1.0	0.0
19	125.976468	2	737.664	0.0	1.0
20	144.464165	0	622.000	0.0	1.0
21	185.806000	2	560.000	0.0	0.0
22	60.108241	0	355.850	1.0	0.0
23	71.442407	0	196.095	0.0	0.0
24	79.432065	1	470.250	0.0	1.0
25	83.612700	1	509.400	1.0	0.0
26	42.363768	1	151.848	0.0	0.0
27	62.152107	0	347.880	0.0	1.0
28	176.422797	2	1044.450	1.0	0.0
29	58.807599	0	212.055	0.0	0.0

	SquareFeet	Density	price	CityA	CityB
30	82.683670	1	400.500	0.0	1.0
31	87.886238	0	272.448	0.0	0.0
32	114.735205	0	679.250	1.0	0.0

split train and test data

In [7]:
```
from sklearn.model_selection import train_test_split

x = data[["SquareFeet","Density","CityA","CityB"]]
y = data["price"]

x_train, x_test, y_train, y_test = train_test_split(x,y,test_size=0.2,random_state=
len(x),len(x_train),len(x_test)
```

Out[7]:
```
(33, 26, 7)
```

In [8]:
```
x_train = x_train.to_numpy()
x_test = x_test.to_numpy()
```

In [9]:
```
#y_pred = w1 * x1 + w2 * x2 + w3 * x3 ... + b
#y_pred = w1*SquareFeet + w2*Density + w3*CityA + w4*CityB + b
```

In [10]:
```
import numpy as np

w = np.array([80.5,1,1,0])
b = 1

y_pred = (w*x_train).sum(axis = 1) + b
y_pred
```

Out[10]:
```
array([ 7629.26533  , 15461.4553305,  3412.283324 ,  4908.021624 ,
        7406.904585 ,  4379.0345275,  4735.0117195,  4115.280325 ,
        7075.842159 ,  5896.208902 , 14960.383    ,  4319.2049955,
        6358.887775 , 11630.3652825,  6658.035435 ,  6583.24852  ,
        6315.015626 ,  9141.961013 ,  9238.1840025,  5004.2446135,
        5752.1137635,  4840.7134005,  4975.3298475,  5566.146476 ,
        7399.4258935, 10144.105674 ])
```

In [11]:
```
def compute_cost(x,y,w,b):
    y_pred = (w*x).sum(axis = 1) + b
    cost = ((y - y_pred)**2).mean()
    return cost
```

In [12]:
```
w = np.array([80.5,1,1,0])
b = 1
compute_cost(x_train,y_train,w,b)
```

Out[12]:
```
np.float64(52763482.698299974)
```

In [ ]: