DAY 9:

World Dataset

Linear Regression

In [1]:

```
#to import libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [2]:

```
df=pd.read_csv(r"C:\Users\user\Downloads\18_world-data-2023.csv")[0:50]
df
```

Out[2]:

	Country	Density\n(P/Km2)	Abbreviation	Agricultural Land(%)	Land Area(Km2)	Armed Forces size	Birth Rate	Ca (
0	Afghanistan	60	AF	58.10%	652,230	323,000	32.49	
1	Albania	105	AL	43.10%	28,748	9,000	11.78	3
2	Algeria	18	DZ	17.40%	2,381,741	317,000	24.28	2
3	Andorra	164	AD	40.00%	468	NaN	7.20	3
4	Angola	26	AO	47.50%	1,246,700	117,000	40.73	2
5	Antigua and Barbuda	223	AG	20.50%	443	0	15.33	
6	Argentina	17	AR	54.30%	2,780,400	105,000	17.02	
7	Armenia	104	AM	58.90%	29,743	49,000	13.99	3
8	Australia	3	AU	48.20%	7,741,220	58,000	12.60	
9	Austria	109	AT	32.40%	83,871	21,000	9.70	
10	Azerbaijan	123	AZ	57.70%	86,600	82,000	14.00	9
11	The Bahamas	39	BS	1.40%	13,880	1,000	13.97	
12	Bahrain	2,239	BH	11.10%	765	19,000	13.99	9
13	Bangladesh	1,265	BD	70.60%	148,460	221,000	18.18	8
14	Barbados	668	BB	23.30%	430	1,000	10.65	
15	Belarus	47	BY	42.00%	207,600	155,000	9.90	3
16	Belgium	383	BE	44.60%	30,528	32,000	10.30	
17	Belize	17	BZ	7.00%	22,966	2,000	20.79	5
18	Benin	108	BJ	33.30%	112,622	12,000	36.22	2
19	Bhutan	20	BT	13.60%	38,394	6,000	17.26	9
20	Bolivia	11	BO	34.80%	1,098,581	71,000	21.75	5
21	Bosnia and Herzegovina	64	BA	43.10%	51,197	11,000	8.11	3
22	Botswana	4	BW	45.60%	581,730	9,000	24.82	2
23	Brazil	25	BR	33.90%	8,515,770	730,000	13.92	
24	Brunei	83	BN	2.70%	5,765	8,000	14.90	6
25	Bulgaria	64	BG	46.30%	110,879	31,000	8.90	3
26	Burkina Faso	76	BF	44.20%	274,200	11,000	37.93	2
27	Burundi	463	BI	79.20%	27,830	31,000	39.01	2
28	Ivory Coast	83	CI	64.80%	322,463	27,000	35.74	2
29	Cape Verde	138	CV	19.60%	4,033	1,000	19.49	2
30	Cambodia	95	KH	30.90%	181,035	191,000	22.46	8
31	Cameroon	56	CM	20.60%	475,440	24,000	35.39	2
32	Canada	4	CA	6.90%	9,984,670	72,000	10.10	

	Country	Density\n(P/Km2)	Abbreviation	Agricultural Land(%)	Land Area(Km2)	Armed Forces size	Birth Rate	Ca (
33	Central African Republic	8	CF	8.20%	622,984	8,000	35.35	2
34	Chad	13	TD	39.70%	1,284,000	35,000	42.17	2
35	Chile	26	CL	21.20%	756,096	122,000	12.43	
36	China	153	CN	56.20%	9,596,960	2,695,000	10.90	
37	Colombia	46	CO	40.30%	1,138,910	481,000	14.88	
38	Comoros	467	KM	71.50%	2,235	NaN	31.88	2
39	Republic of the Congo	16	NaN	31.10%	342,000	12,000	32.86	2
40	Costa Rica	100	CR	34.50%	51,100	10,000	13.97	5
41	Croatia	73	HR	27.60%	56,594	18,000	9.00	3
42	Cuba	106	CU	59.90%	110,860	76,000	10.17	
43	Cyprus	131	CY	12.20%	9,251	16,000	10.46	3
44	Czech Republic	139	CZ	45.20%	78,867	23,000	10.70	4
45	Democratic Republic of the Congo	40	CD	11.60%	2,344,858	134,000	41.18	2
46	Denmark	137	DK	62.00%	43,094	15,000	10.60	
47	Djibouti	43	DJ	73.40%	23,200	13,000	21.47	2
48	Dominica	96	DM	33.30%	751	NaN	12.00	
49	Dominican Republic	225	DO	48.70%	48,670	71,000	19.51	

50 rows × 35 columns

In [3]:

```
df.head()
```

Out[3]:

	Country	Density\n(P/Km2)	Abbreviation	Agricultural Land(%)	Land Area(Km2)	Armed Forces size	Birth Rate	Callin Cod
0	Afghanistan	60	AF	58.10%	652,230	323,000	32.49	93.
1	Albania	105	AL	43.10%	28,748	9,000	11.78	355.
2	Algeria	18	DZ	17.40%	2,381,741	317,000	24.28	213.
3	Andorra	164	AD	40.00%	468	NaN	7.20	376.
4	Angola	26	AO	47.50%	1,246,700	117,000	40.73	244.

5 rows × 35 columns

In [4]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 50 entries, 0 to 49
Data columns (total 35 columns):
 #   Column                                      Non-Null Count  Dtype  
---  ------                                      --------------  -----  
 0   Country                                     50 non-null     object 
 1   Density
(P/Km2)                          50 non-null     object 
 2   Abbreviation                                49 non-null     object 
 3   Agricultural Land( %)                       50 non-null     object 
 4   Land Area(Km2)                              50 non-null     object 
 5   Armed Forces size                           47 non-null     object 
 6   Birth Rate                                  50 non-null     float64
 7   Calling Code                                50 non-null     float64
 8   Capital/Major City                          50 non-null     object 
 9   Co2-Emissions                               50 non-null     object 
 10  CPI                                         47 non-null     object 
 11  CPI Change (%)                              48 non-null     object 
 12  Currency-Code                               46 non-null     object 
 13  Fertility Rate                              50 non-null     float64
 14  Forested Area (%)                           50 non-null     object 
 15  Gasoline Price                              48 non-null     object 
 16  GDP                                         50 non-null     object 
 17  Gross primary education enrollment (%)      49 non-null     object 
 18  Gross tertiary education enrollment (%)     48 non-null     object 
 19  Infant mortality                            50 non-null     float64
 20  Largest city                                49 non-null     object 
 21  Life expectancy                             49 non-null     float64
 22  Maternal mortality ratio                    48 non-null     float64
 23  Minimum wage                                42 non-null     object 
 24  Official language                           50 non-null     object 
 25  Out of pocket health expenditure            49 non-null     object 
 26  Physicians per thousand                     50 non-null     float64
 27  Population                                  50 non-null     object 
 28  Population: Labor force participation (%)   47 non-null     object 
 29  Tax revenue (%)                             44 non-null     object 
 30  Total tax rate                              48 non-null     object 
 31  Unemployment rate                           47 non-null     object 
 32  Urban_population                            50 non-null     object 
 33  Latitude                                    50 non-null     float64
 34  Longitude                                   50 non-null     float64
dtypes: float64(9), object(26)
memory usage: 13.8+ KB
```

In [5]:

```
#to display summary of statistics
df.describe()
```

Out[5]:

	Birth Rate	Calling Code	Fertility Rate	Infant mortality	Life expectancy	Maternal mortality ratio	Physicians per thousand	La
count	50.00000	50.000000	50.00000	50.000000	49.000000	48.000000	50.000000	50.0
mean	19.64860	291.820000	2.62600	22.618000	72.312245	174.041667	1.929800	17.6
std	10.67511	272.353663	1.41232	22.042368	7.988498	248.707549	1.782451	24.0
min	7.20000	1.000000	1.27000	1.900000	52.800000	2.000000	0.040000	-38.4
25%	10.75000	56.250000	1.66000	5.225000	66.600000	13.750000	0.272500	5.0
50%	14.89000	240.000000	1.94000	11.750000	74.900000	50.000000	1.665000	16.7
75%	24.68500	375.750000	2.98250	35.375000	78.100000	242.750000	2.972500	39.0
max	42.17000	994.000000	5.92000	84.500000	82.700000	1140.000000	8.420000	56.2

In [6]:

```
#to display cloumn heading
df.columns
```

Out[6]:

```
Index(['Country', 'Density\n(P/Km2)', 'Abbreviation', 'Agricultural Land(
%)',
       'Land Area(Km2)', 'Armed Forces size', 'Birth Rate', 'Calling Cod
e',
       'Capital/Major City', 'Co2-Emissions', 'CPI', 'CPI Change (%)',
       'Currency-Code', 'Fertility Rate', 'Forested Area (%)',
       'Gasoline Price', 'GDP', 'Gross primary education enrollment (%)',
       'Gross tertiary education enrollment (%)', 'Infant mortality',
       'Largest city', 'Life expectancy', 'Maternal mortality ratio',
       'Minimum wage', 'Official language', 'Out of pocket health expendit
ure',
       'Physicians per thousand', 'Population',
       'Population: Labor force participation (%)', 'Tax revenue (%)',
       'Total tax rate', 'Unemployment rate', 'Urban_population', 'Latitud
e',
       'Longitude'],
      dtype='object')
```

EDA and VISUALIZATION

In [7]:

```
df1=df[['Birth Rate', 'Calling Code','Fertility Rate', 'Infant mortality','Life expectanc
       'Longitude']]
df1
```

Out[7]:

	Birth Rate	Calling Code	Fertility Rate	Infant mortality	Life expectancy	Maternal mortality ratio	Latitude	Longitude
0	32.49	93.0	4.47	47.9	64.5	638.0	33.939110	67.709953
1	11.78	355.0	1.62	7.8	78.5	15.0	41.153332	20.168331
2	24.28	213.0	3.02	20.1	76.7	112.0	28.033886	1.659626
3	7.20	376.0	1.27	2.7	NaN	NaN	42.506285	1.521801
4	40.73	244.0	5.52	51.6	60.8	241.0	-11.202692	17.873887
5	15.33	1.0	1.99	5.0	76.9	42.0	17.060816	-61.796428
6	17.02	54.0	2.26	8.8	76.5	39.0	-38.416097	-63.616672
7	13.99	374.0	1.76	11.0	74.9	26.0	40.069099	45.038189
8	12.60	61.0	1.74	3.1	82.7	6.0	-25.274398	133.775136
9	9.70	43.0	1.47	2.9	81.6	5.0	47.516231	14.550072
10	14.00	994.0	1.73	19.2	72.9	26.0	40.143105	47.576927
11	13.97	1.0	1.75	8.3	73.8	70.0	25.034280	-77.396280
12	13.99	973.0	1.99	6.1	77.2	14.0	26.066700	50.557700
13	18.18	880.0	2.04	25.1	72.3	173.0	23.684994	90.356331
14	10.65	1.0	1.62	11.3	79.1	27.0	13.193887	-59.543198
15	9.90	375.0	1.45	2.6	74.2	2.0	53.709807	27.953389
16	10.30	32.0	1.62	2.9	81.6	5.0	50.503887	4.469936
17	20.79	501.0	2.31	11.2	74.5	36.0	17.189877	-88.497650
18	36.22	229.0	4.84	60.5	61.5	397.0	9.307690	2.315834
19	17.26	975.0	1.98	24.8	71.5	183.0	27.514162	90.433601
20	21.75	591.0	2.73	21.8	71.2	155.0	-16.290154	-63.588653
21	8.11	387.0	1.27	5.0	77.3	10.0	43.915886	17.679076
22	24.82	267.0	2.87	30.0	69.3	144.0	-22.328474	24.684866
23	13.92	55.0	1.73	12.8	75.7	60.0	-14.235004	-51.925280
24	14.90	673.0	1.85	9.8	75.7	31.0	4.535277	114.727669
25	8.90	359.0	1.56	5.9	74.9	10.0	42.733883	25.485830
26	37.93	226.0	5.19	49.0	61.2	320.0	12.238333	-1.561593
27	39.01	257.0	5.41	41.0	61.2	548.0	-3.373056	29.918886
28	35.74	225.0	4.65	59.4	57.4	617.0	7.539989	-5.547080
29	19.49	238.0	2.27	16.7	72.8	58.0	16.538800	-23.041800
30	22.46	855.0	2.50	24.0	69.6	160.0	12.565679	104.990963
31	35.39	237.0	4.57	50.6	58.9	529.0	7.369722	12.354722
32	10.10	1.0	1.50	4.3	81.9	10.0	56.130366	-106.346771
33	35.35	236.0	4.72	84.5	52.8	829.0	6.611111	20.939444
34	42.17	235.0	5.75	71.4	54.0	1140.0	15.454166	18.732207
35	12.43	56.0	1.65	6.2	80.0	13.0	-35.675147	-71.542969

	Birth Rate	Calling Code	Fertility Rate	Infant mortality	Life expectancy	Maternal mortality ratio	Latitude	Longitude
36	10.90	86.0	1.69	7.4	77.0	29.0	35.861660	104.195397
37	14.88	57.0	1.81	12.2	77.1	83.0	4.570868	-74.297333
38	31.88	269.0	4.21	51.3	64.1	273.0	-11.645500	43.333300
39	32.86	242.0	4.43	36.2	64.3	378.0	-0.228021	15.827659
40	13.97	506.0	1.75	7.6	80.1	27.0	9.748917	-83.753428
41	9.00	385.0	1.47	4.0	78.1	8.0	45.100000	15.200000
42	10.17	53.0	1.62	3.7	78.7	36.0	21.521757	-77.781167
43	10.46	357.0	1.33	1.9	80.8	6.0	35.126413	33.429859
44	10.70	420.0	1.69	2.7	79.0	3.0	49.817492	15.472962
45	41.18	243.0	5.92	68.2	60.4	473.0	-4.038333	21.758664
46	10.60	45.0	1.73	3.6	81.0	4.0	56.263920	9.501785
47	21.47	253.0	2.73	49.8	66.6	248.0	11.825138	42.590275
48	12.00	1.0	1.90	32.9	76.6	NaN	15.414999	-61.370976
49	19.51	1.0	2.35	24.1	73.9	95.0	18.735693	-70.162651

In [8]:

```
df1.info()
```

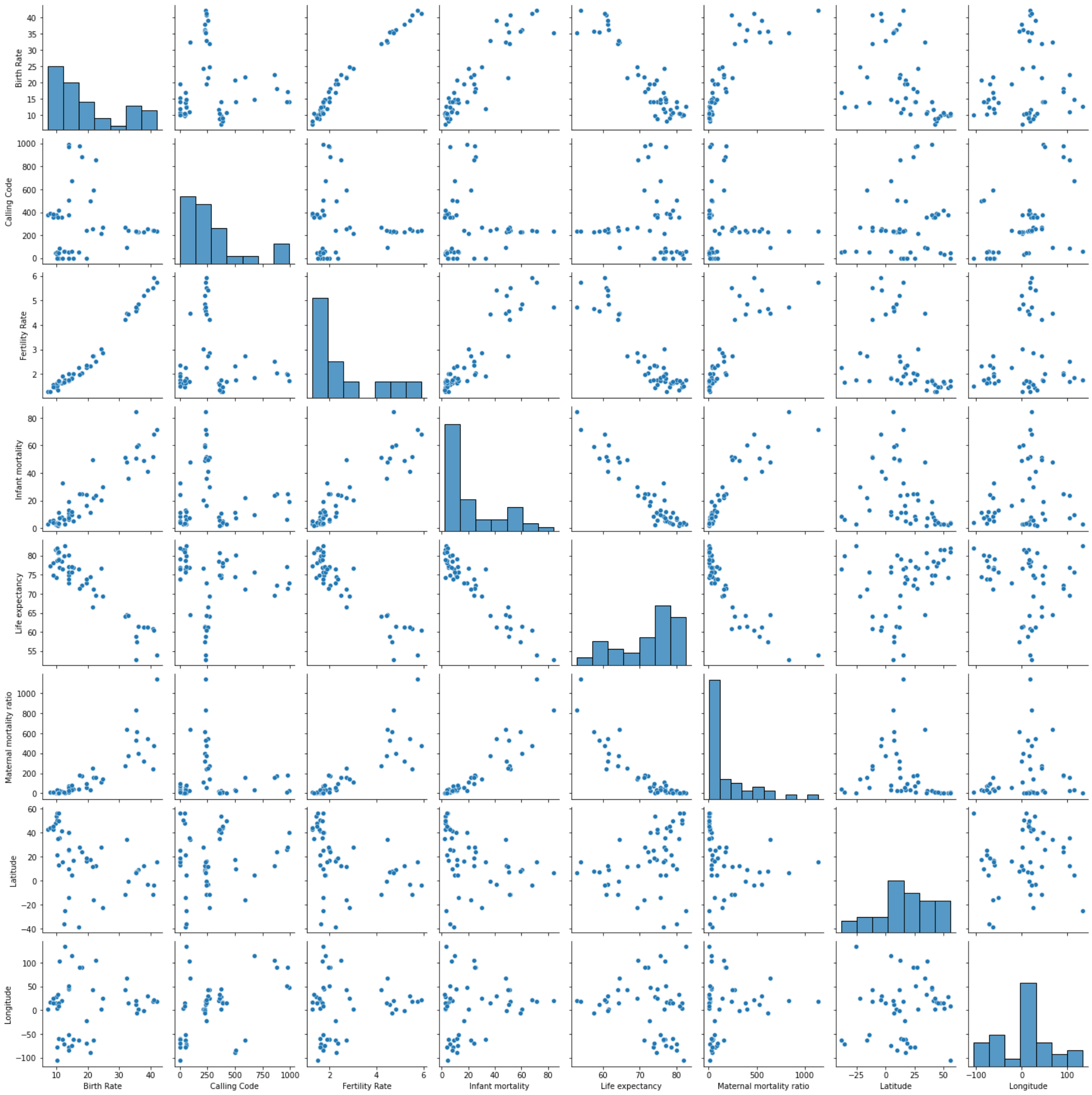
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 50 entries, 0 to 49
Data columns (total 8 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   Birth Rate                50 non-null     float64
 1   Calling Code              50 non-null     float64
 2   Fertility Rate            50 non-null     float64
 3   Infant mortality          50 non-null     float64
 4   Life expectancy           49 non-null     float64
 5   Maternal mortality ratio  48 non-null     float64
 6   Latitude                  50 non-null     float64
 7   Longitude                 50 non-null     float64
dtypes: float64(8)
memory usage: 3.2 KB
```

In [9]:

```
sns.pairplot(df1)
```

Out[9]:

```
<seaborn.axisgrid.PairGrid at 0x21fd3684be0>
```



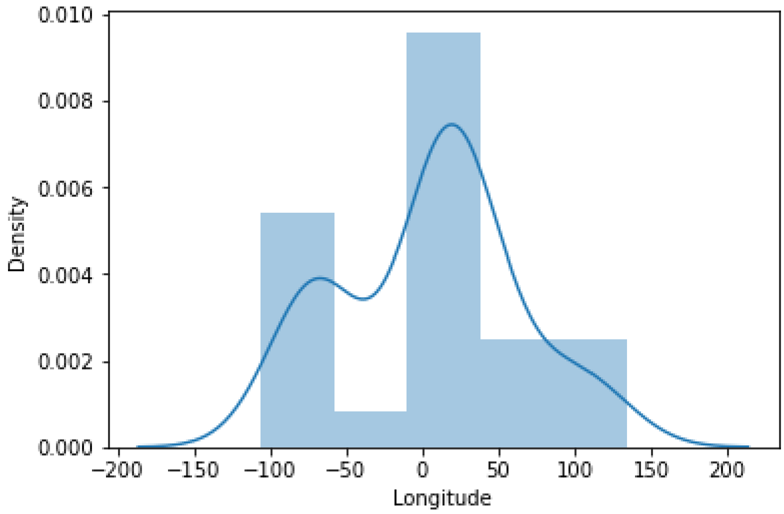
In [10]:

```
sns.distplot(df['Longitude'])
```

```
C:\ProgramData\Anaconda3\lib\site-packages\seaborn\distributions.py:2557:
FutureWarning: `distplot` is a deprecated function and will be removed in
a future version. Please adapt your code to use either `displot` (a figure
-level function with similar flexibility) or `histplot` (an axes-level fun
ction for histograms).
  warnings.warn(msg, FutureWarning)
```

Out[10]:

```
<AxesSubplot:xlabel='Longitude', ylabel='Density'>
```

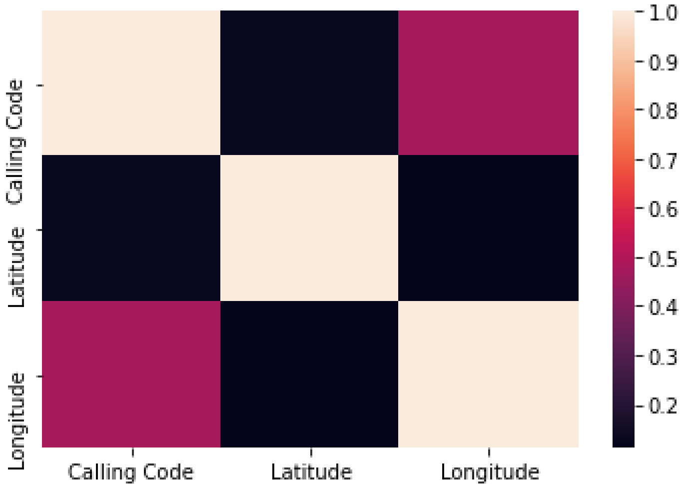


In [11]:

```
data=df1[['Calling Code','Latitude',
       'Longitude']]
sns.heatmap(data.corr())
```

Out[11]:

```
<AxesSubplot:>
```



to Train the model-Model buliding

we are going to split our data into two variable where x is a independent and y is dependent on x

In [12]:

```
x=data[['Calling Code','Latitude']]
y=data['Longitude']
```

In [13]:

```
# to split my dataset into test and train 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 [14]:

```
from sklearn.linear_model import LinearRegression

lr=LinearRegression()
lr.fit(x_train,y_train)
```

Out[14]:

```
LinearRegression()
```

In [15]:

```
print(lr.intercept_)
```

```
-11.770187454909436
```

In [16]:

```
coeff=pd.DataFrame(lr.coef_,x.columns,columns=['Co-effecient'])
coeff
```

Out[16]:

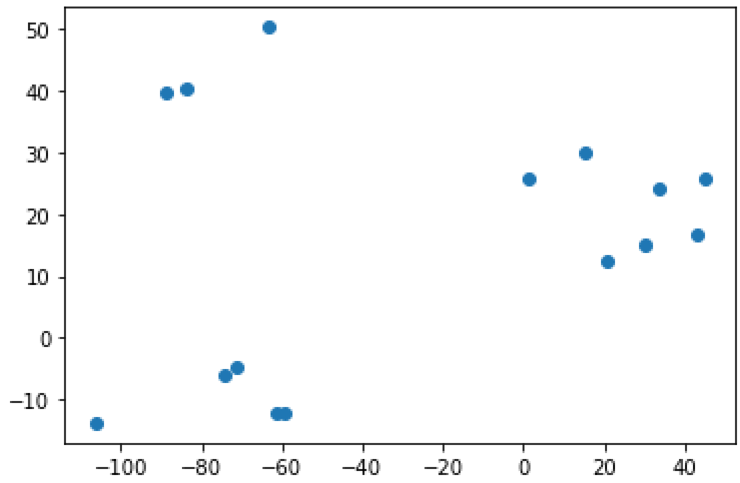
	Co-effecient
Calling Code	0.104014
Latitude	-0.037939

In [17]:

```
prediction=lr.predict(x_test)
plt.scatter(y_test,prediction)
```

Out[17]:

```
<matplotlib.collections.PathCollection at 0x21fd7c326d0>
```



In [18]:

```
print(lr.score(x_test,y_test))
```

```
-0.6146745669577893
```

In [19]:

```
lr.score(x_train,y_train)
```

Out[19]:

```
0.32441522905909725
```

Ridge Regression

In [20]:

```
from sklearn.linear_model import Ridge,Lasso
```

In [21]:

```
rr=Ridge(alpha=10)
rr.fit(x_train,y_train)
rr.score(x_test,y_test)
```

Out[21]:

```
-0.6146693807757213
```

Lasso Regression

In [22]:

```
la=Lasso(alpha=10)
la.fit(x_train,y_train)
```

Out[22]:

```
Lasso(alpha=10)
```

In [23]:

```
la.score(x_test,y_test)
```

Out[23]:

```
-0.6100411691130756
```

In [ ]:

```

```