In [1]:

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

14.IRIS

In [3]:

```
a=pd.read_csv(r"C:\Users\user\Downloads\14_Iris.csv")
a
```

Out[3]:

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
0	1	5.1	3.5	1.4	0.2	Iris-setosa
1	2	4.9	3.0	1.4	0.2	Iris-setosa
2	3	4.7	3.2	1.3	0.2	Iris-setosa
3	4	4.6	3.1	1.5	0.2	Iris-setosa
4	5	5.0	3.6	1.4	0.2	Iris-setosa
...	...	...	...	...	...	...
145	146	6.7	3.0	5.2	2.3	Iris-virginica
146	147	6.3	2.5	5.0	1.9	Iris-virginica
147	148	6.5	3.0	5.2	2.0	Iris-virginica
148	149	6.2	3.4	5.4	2.3	Iris-virginica
149	150	5.9	3.0	5.1	1.8	Iris-virginica

150 rows × 6 columns

In [4]:

```
b=a.head(100)
b
```

Out[4]:

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
0	1	5.1	3.5	1.4	0.2	Iris-setosa
1	2	4.9	3.0	1.4	0.2	Iris-setosa
2	3	4.7	3.2	1.3	0.2	Iris-setosa
3	4	4.6	3.1	1.5	0.2	Iris-setosa
4	5	5.0	3.6	1.4	0.2	Iris-setosa
...	...	...	...	...	...	...
95	96	5.7	3.0	4.2	1.2	Iris-versicolor
96	97	5.7	2.9	4.2	1.3	Iris-versicolor
97	98	6.2	2.9	4.3	1.3	Iris-versicolor
98	99	5.1	2.5	3.0	1.1	Iris-versicolor
99	100	5.7	2.8	4.1	1.3	Iris-versicolor

100 rows × 6 columns

In [5]:

```
b.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100 entries, 0 to 99
Data columns (total 6 columns):
 #   Column         Non-Null Count  Dtype  
---  ------         --------------  -----  
 0   Id             100 non-null    int64  
 1   SepalLengthCm  100 non-null    float64
 2   SepalWidthCm   100 non-null    float64
 3   PetalLengthCm  100 non-null    float64
 4   PetalWidthCm   100 non-null    float64
 5   Species        100 non-null    object 
dtypes: float64(4), int64(1), object(1)
memory usage: 4.8+ KB
```

In [6]:

```
b.describe()
```

Out[6]:

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm
count	100.000000	100.000000	100.000000	100.000000	100.000000
mean	50.500000	5.471000	3.094000	2.862000	0.785000
std	29.011492	0.641698	0.476057	1.448565	0.566288
min	1.000000	4.300000	2.000000	1.000000	0.100000
25%	25.750000	5.000000	2.800000	1.500000	0.200000
50%	50.500000	5.400000	3.050000	2.450000	0.800000
75%	75.250000	5.900000	3.400000	4.325000	1.300000
max	100.000000	7.000000	4.400000	5.100000	1.800000

In [7]:

```
b.columns
```

Out[7]:

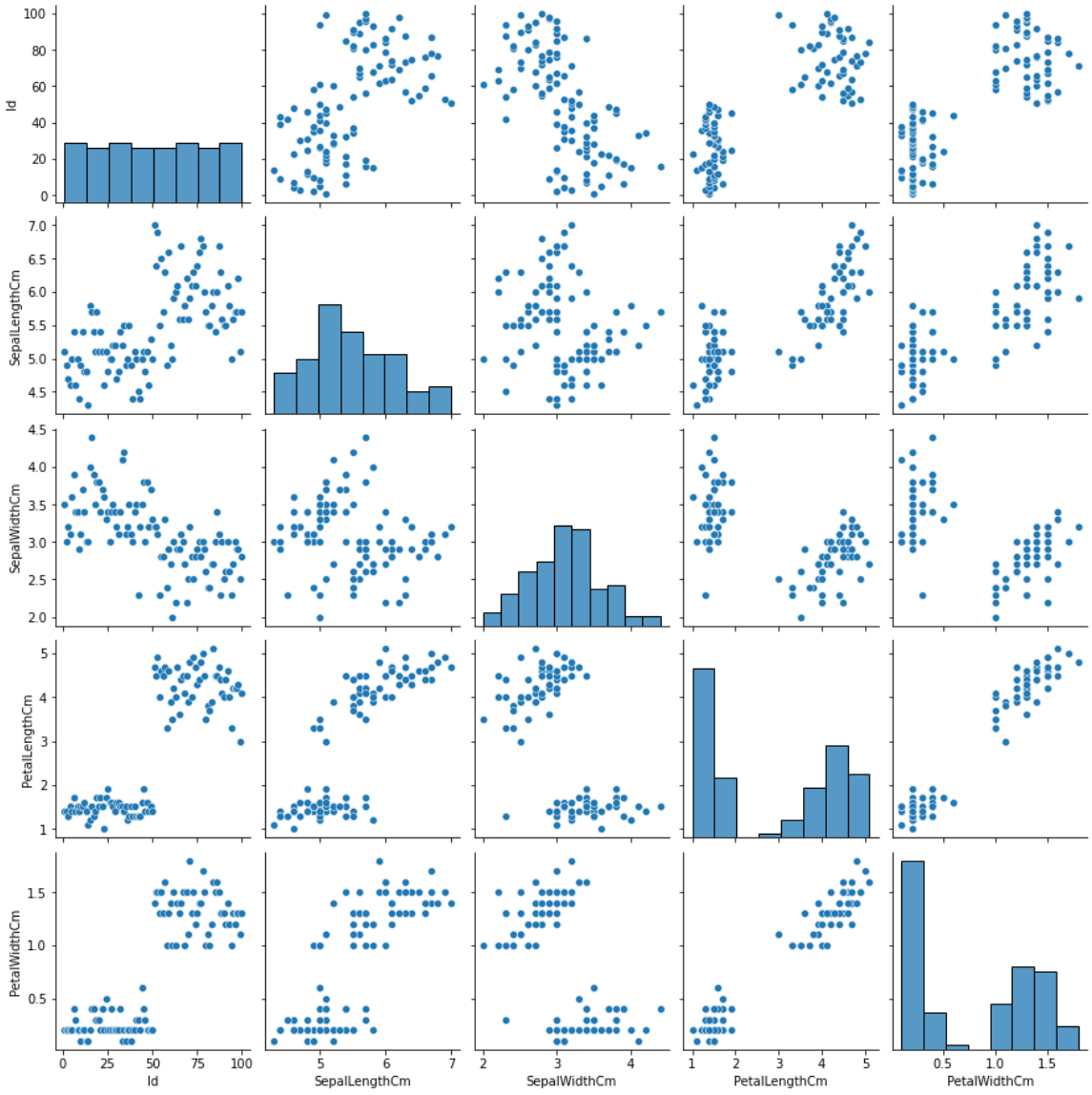
```
Index(['Id', 'SepalLengthCm', 'SepalWidthCm', 'PetalLengthCm', 'PetalWidth
Cm',
       'Species'],
      dtype='object')
```

In [8]:

```
sns.pairplot(b)
```

Out[8]:

```
<seaborn.axisgrid.PairGrid at 0x1b03cca61f0>
```



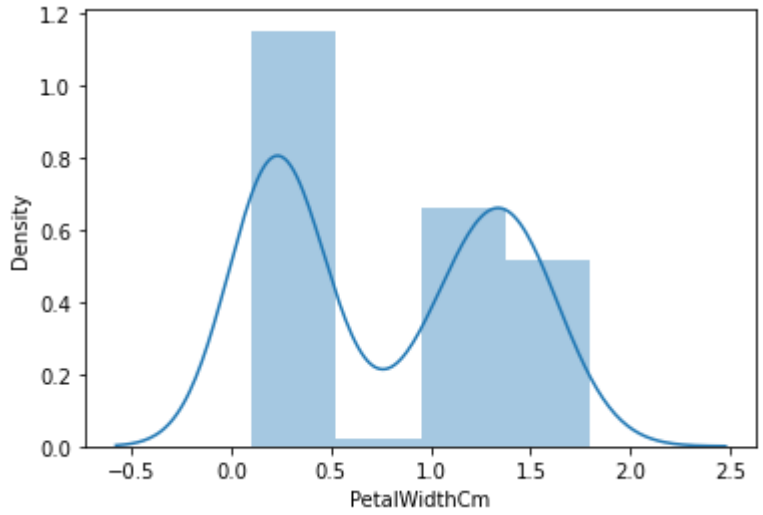
In [10]:

```
sns.distplot(b['PetalWidthCm'])
```

```
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='PetalWidthCm', ylabel='Density'>
```



In [11]:

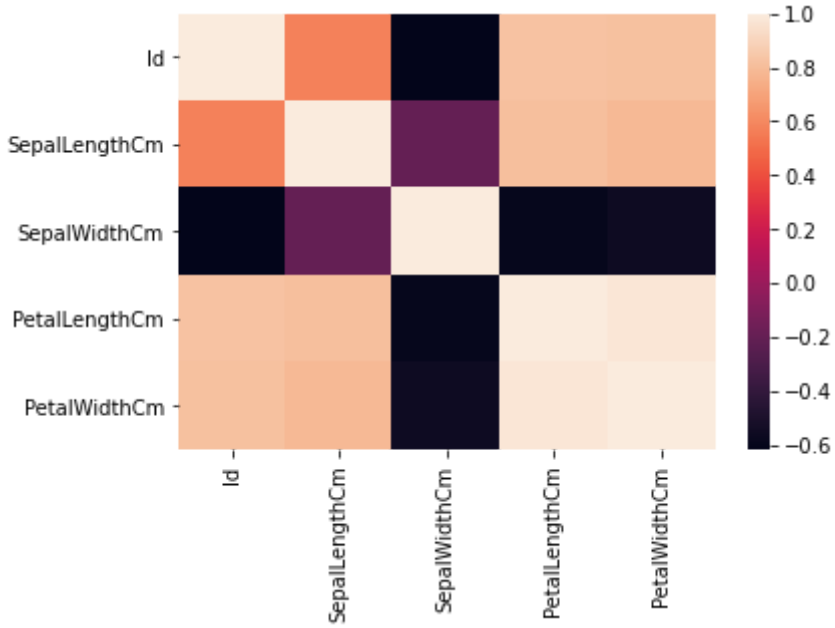
```
f=b[['Id', 'SepalLengthCm', 'SepalWidthCm', 'PetalLengthCm', 'PetalWidthCm',
       'Species']]
```

In [12]:

```
sns.heatmap(f.corr())
```

Out[12]:

<AxesSubplot:>



In [17]:

```
x=f[['Id', 'SepalLengthCm', 'SepalWidthCm', 'PetalLengthCm']]
y=f['PetalWidthCm']
```

In [18]:

```
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.5)
```

In [19]:

```
from sklearn.linear_model import LinearRegression

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

Out[19]:

```
LinearRegression()
```

In [20]:

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

```
-0.45844839820297834
```

In [21]:

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

Out[21]:

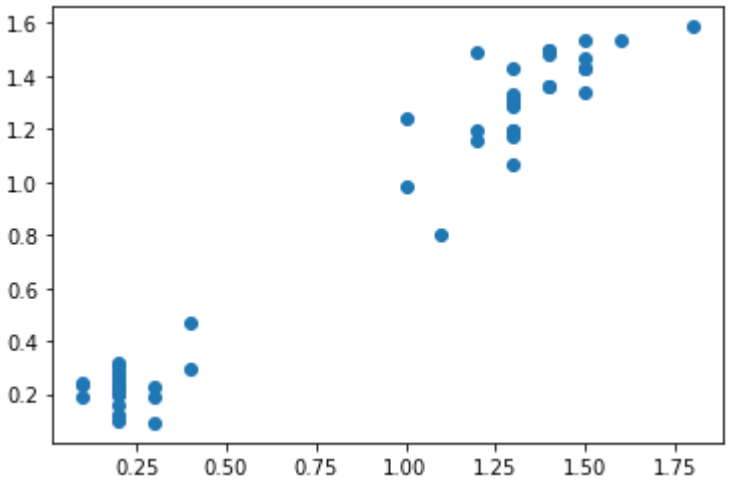
	Co-efficient
Id	0.000047
SepalLengthCm	-0.054424
SepalWidthCm	0.107123
PetalLengthCm	0.420683

In [22]:

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

Out[22]:

```
<matplotlib.collections.PathCollection at 0x1b043cc8610>
```



In [23]:

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

```
0.9604252477594315
```

In [24]:

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

Out[24]:

```
0.9577784111815976
```

RIDGE REGRESSION

In [25]:

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

In [26]:

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

Out[26]:

```
Ridge(alpha=10)
```

In [27]:

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

Out[27]:

```
0.9509851959606976
```

LASSO REGRESSION

In [28]:

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

Out[28]:

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

In [29]:

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

Out[29]:

```
0.22658608754030318
```

15.Horse Racing Results

In [35]:

```
a=pd.read_csv(r"C:\Users\user\Downloads\15_Horse Racing Results.csv - 15_Horse Racing Res
a
```

Out[35]:

	Dato	Track	Race Number	Distance	Surface	Prize money	Starting position	Jockey	Jockey weight	C
0	03.09.2017	Sha Tin	10	1400	Gress	1310000	6	K C Leung	52	
1	16.09.2017	Sha Tin	10	1400	Gress	1310000	14	C Y Ho	52	
2	14.10.2017	Sha Tin	10	1400	Gress	1310000	8	C Y Ho	52	
3	11.11.2017	Sha Tin	9	1600	Gress	1310000	13	Brett Prebble	54	
4	26.11.2017	Sha Tin	9	1600	Gress	1310000	9	C Y Ho	52	
...	...	...	...	...	...	...	...	...	...	
27003	14.06.2020	Sha Tin	11	1200	Gress	1450000	6	A Hamelin	59	A
27004	21.06.2020	Sha Tin	2	1200	Gress	967000	7	K C Leung	57	A
27005	21.06.2020	Sha Tin	4	1200	Gress	967000	6	Blake Shinn	57	A
27006	21.06.2020	Sha Tin	5	1200	Gress	967000	14	Joao Moreira	57	Z
27007	21.06.2020	Sha Tin	11	1200	Gress	1450000	7	C Schofield	55	Z

27008 rows × 21 columns

In [36]:

```
b=a.head(100)
b
```

Out[36]:

	Dato	Track	Race Number	Distance	Surface	Prize money	Starting position	Jockey	Jockey weight	Cou
0	03.09.2017	Sha Tin	10	1400	Gress	1310000	6	K C Leung	52	Sv
1	16.09.2017	Sha Tin	10	1400	Gress	1310000	14	C Y Ho	52	Sv
2	14.10.2017	Sha Tin	10	1400	Gress	1310000	8	C Y Ho	52	Sv
3	11.11.2017	Sha Tin	9	1600	Gress	1310000	13	Brett Prebble	54	Sv
4	26.11.2017	Sha Tin	9	1600	Gress	1310000	9	C Y Ho	52	Sv
...	...	...	...	...	...	...	...	...	...	
95	10.12.2017	Sha Tin	5	1200	Gress	18500000	13	Francois-Xavier Bertras	57	(B
96	10.12.2017	Sha Tin	7	1600	Gress	23000000	11	Ryan Moore	57	
97	01.10.2017	Sha Tin	7	1000	Gress	3000000	10	Brett Prebble	59	Zea
98	22.10.2017	Sha Tin	7	1200	Gress	4000000	9	Brett Prebble	59	Zea
99	19.11.2017	Sha Tin	7	1200	Gress	4000000	3	Brett Prebble	56	Zea

100 rows × 21 columns

In [37]:

```
b.describe()
```

Out[37]:

	Race Number	Distance	Prize money	Starting position	Jockey weight	Horse age	Path
count	100.000000	100.000000	1.000000e+02	100.000000	100.000000	100.00000	100.000000
mean	6.910000	1446.000000	3.562200e+06	6.170000	55.870000	6.58000	1.510000
std	2.099038	334.820923	4.486259e+06	3.440857	2.942736	1.35721	1.573101
min	1.000000	1000.000000	9.200000e+05	1.000000	49.000000	3.00000	0.000000
25%	6.000000	1200.000000	1.380000e+06	3.000000	54.000000	6.00000	0.000000
50%	7.000000	1400.000000	1.950000e+06	6.000000	56.000000	7.00000	1.000000
75%	8.000000	1650.000000	3.000000e+06	9.000000	58.000000	8.00000	3.000000
max	10.000000	2400.000000	2.300000e+07	14.000000	60.000000	9.00000	6.000000

In [38]:

```
b.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100 entries, 0 to 99
Data columns (total 21 columns):
 #   Column             Non-Null Count  Dtype 
---  ------             --------------  ----- 
 0   Dato               100 non-null    object
 1   Track              100 non-null    object
 2   Race Number        100 non-null    int64 
 3   Distance           100 non-null    int64 
 4   Surface            100 non-null    object
 5   Prize money        100 non-null    int64 
 6   Starting position  100 non-null    int64 
 7   Jockey             100 non-null    object
 8   Jockey weight      100 non-null    int64 
 9   Country            100 non-null    object
 10  Horse age          100 non-null    int64 
 11  TrainerName        100 non-null    object
 12  Race time          100 non-null    object
 13  Path               100 non-null    int64 
 14  Final place        100 non-null    int64 
 15  FGrating           100 non-null    int64 
 16  Odds               100 non-null    object
 17  RaceType           100 non-null    object
 18  HorseId            100 non-null    int64 
 19  JockeyId           100 non-null    int64 
 20  TrainerID          100 non-null    int64 
dtypes: int64(12), object(9)
memory usage: 16.5+ KB
```

In [39]:

```
c=b.dropna(axis=1)
c
```

Out[39]:

	Dato	Track	Race Number	Distance	Surface	Prize money	Starting position	Jockey	Jockey weight	Cou
0	03.09.2017	Sha Tin	10	1400	Gress	1310000	6	K C Leung	52	Sv
1	16.09.2017	Sha Tin	10	1400	Gress	1310000	14	C Y Ho	52	Sv
2	14.10.2017	Sha Tin	10	1400	Gress	1310000	8	C Y Ho	52	Sv
3	11.11.2017	Sha Tin	9	1600	Gress	1310000	13	Brett Prebble	54	Sv
4	26.11.2017	Sha Tin	9	1600	Gress	1310000	9	C Y Ho	52	Sv
...	...	...	...	...	...	...	...	...	...	
95	10.12.2017	Sha Tin	5	1200	Gress	18500000	13	Francois-Xavier Bertras	57	(B
96	10.12.2017	Sha Tin	7	1600	Gress	23000000	11	Ryan Moore	57	
97	01.10.2017	Sha Tin	7	1000	Gress	3000000	10	Brett Prebble	59	Zea
98	22.10.2017	Sha Tin	7	1200	Gress	4000000	9	Brett Prebble	59	Zea
99	19.11.2017	Sha Tin	7	1200	Gress	4000000	3	Brett Prebble	56	Zea

100 rows × 21 columns

In [40]:

```
c.columns
```

Out[40]:

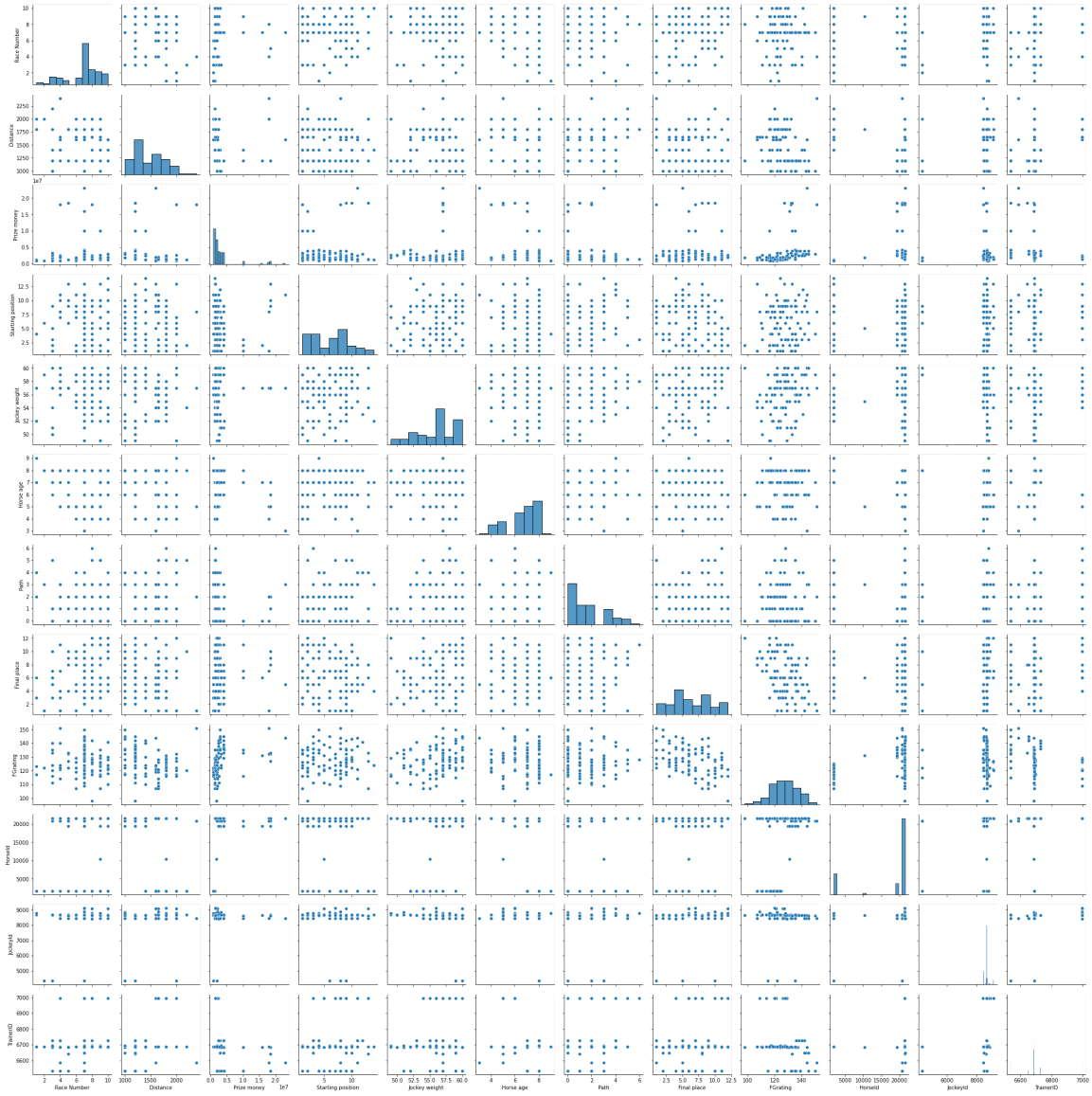
```
Index(['Dato', 'Track', 'Race Number', 'Distance', 'Surface', 'Prize mone
y',
       'Starting position', 'Jockey', 'Jockey weight', 'Country', 'Horse a
ge',
       'TrainerName', 'Race time', 'Path', 'Final place', 'FGrating', 'Odd
s',
       'RaceType', 'HorseId', 'JockeyId', 'TrainerID'],
      dtype='object')
```

In [41]:

```
sns.pairplot(c)
```

Out[41]:

```
<seaborn.axisgrid.PairGrid at 0x1b043cf2e80>
```



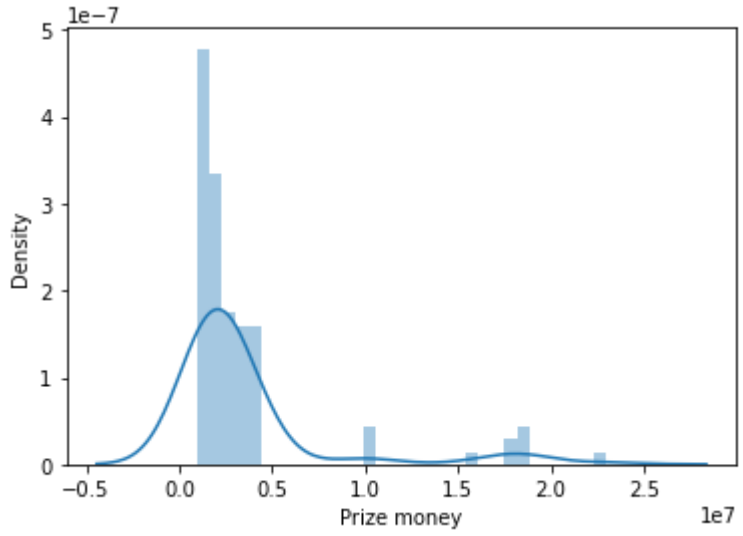
In [42]:

```
sns.distplot(c['Prize money'])
```

```
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[42]:

```
<AxesSubplot:xlabel='Prize money', ylabel='Density'>
```



In [43]:

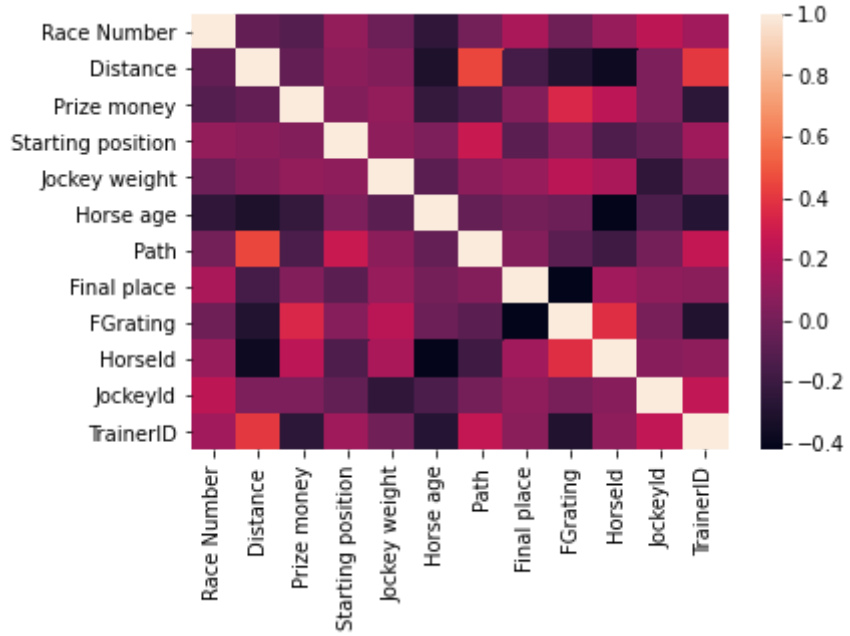
```
f=c[['Dato','Race Number', 'Distance','Prize money',
       'Starting position','Jockey weight','Horse age',
      'Race time', 'Path', 'Final place', 'FGrating', 'Odds',
       'HorseId', 'JockeyId', 'TrainerID']]
```

In [44]:

```
sns.heatmap(f.corr())
```

Out[44]:

<AxesSubplot:>



In [56]:

```
x=f[[ 'Distance','Starting position','Jockey weight','Horse age','FGrating','TrainerID']]
y=f['Prize money']
```

In [57]:

```
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.5)
```

In [58]:

```
from sklearn.linear_model import LinearRegression

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

Out[58]:

```
LinearRegression()
```

In [59]:

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

```
8175909.272110213
```

In [60]:

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

Out[60]:

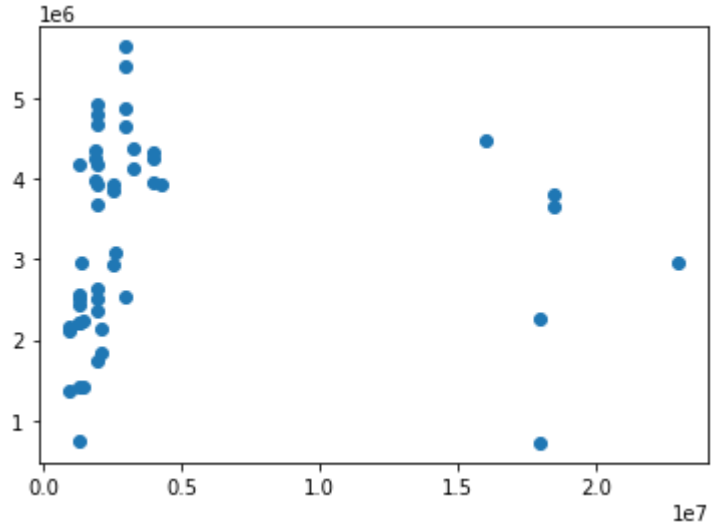
	Co-efficient
Distance	-3147.219403
Starting position	-94581.551137
Jockey weight	104698.157277
Horse age	-63694.881410
FGrating	18062.682362
TrainerID	-1141.421418

In [61]:

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

Out[61]:

```
<matplotlib.collections.PathCollection at 0x1b04bd9afd0>
```



In [62]:

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

```
-0.07782092347579272
```

In [63]:

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

Out[63]:

```
0.17919787489759276
```

In [64]:

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

In [65]:

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

Out[65]:

```
Ridge(alpha=10)
```

In [66]:

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

Out[66]:

```
-0.07825815658097168
```

In [67]:

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

Out[67]:

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

In [68]:

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

Out[68]:

```
-0.07782193863393116
```

16_Sleep_health_and_lifestyle_dataset

In [69]:

```
a=pd.read_csv(r"C:\Users\user\Downloads\16_Sleep_health_and_lifestyle_dataset.csv")
```

In [70]:

```
b=a.head(100)
b
```

Out[70]:

	Person ID	Gender	Age	Occupation	Sleep Duration	Quality of Sleep	Physical Activity Level	Stress Level	BMI Category	I Pre
0	1	Male	27	Software Engineer	6.1	6	42	6	Overweight	1
1	2	Male	28	Doctor	6.2	6	60	8	Normal	1
2	3	Male	28	Doctor	6.2	6	60	8	Normal	1
3	4	Male	28	Sales Representative	5.9	4	30	8	Obese	1
4	5	Male	28	Sales Representative	5.9	4	30	8	Obese	1
...	...	...	...	...	...	...	...	...	...	
95	96	Female	36	Accountant	7.1	8	60	4	Normal	1
96	97	Female	36	Accountant	7.2	8	60	4	Normal	1
97	98	Female	36	Accountant	7.1	8	60	4	Normal	1
98	99	Female	36	Teacher	7.1	8	60	4	Normal	1
99	100	Female	36	Teacher	7.1	8	60	4	Normal	1

100 rows × 13 columns

In [71]:

```
b.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100 entries, 0 to 99
Data columns (total 13 columns):
 #   Column                   Non-Null Count  Dtype  
---  ------                   --------------  -----  
 0   Person ID                100 non-null    int64  
 1   Gender                   100 non-null    object 
 2   Age                      100 non-null    int64  
 3   Occupation               100 non-null    object 
 4   Sleep Duration           100 non-null    float64
 5   Quality of Sleep         100 non-null    int64  
 6   Physical Activity Level  100 non-null    int64  
 7   Stress Level             100 non-null    int64  
 8   BMI Category             100 non-null    object 
 9   Blood Pressure           100 non-null    object 
 10  Heart Rate               100 non-null    int64  
 11  Daily Steps              100 non-null    int64  
 12  Sleep Disorder           100 non-null    object 
dtypes: float64(1), int64(7), object(5)
memory usage: 10.3+ KB
```

In [72]:

```
b.describe()
```

Out[72]:

	Person ID	Age	Sleep Duration	Quality of Sleep	Physical Activity Level	Stress Level	Heart Rate	
count	100.000000	100.00000	100.000000	100.000000	100.000000	100.000000	100.000000	
mean	50.500000	31.69000	6.871000	6.590000	51.910000	6.420000	71.610000	6
std	29.011492	2.26388	0.766903	1.005992	19.429279	1.485145	4.240009	1
min	1.000000	27.00000	5.800000	4.000000	30.000000	3.000000	65.000000	3
25%	25.750000	30.00000	6.100000	6.000000	30.000000	6.000000	70.000000	5
50%	50.500000	31.50000	7.100000	7.000000	60.000000	6.000000	70.000000	7
75%	75.250000	33.00000	7.700000	7.000000	75.000000	8.000000	72.000000	8
max	100.000000	36.00000	7.900000	8.000000	75.000000	8.000000	85.000000	10

In [75]:

```
b.columns
```

Out[75]:

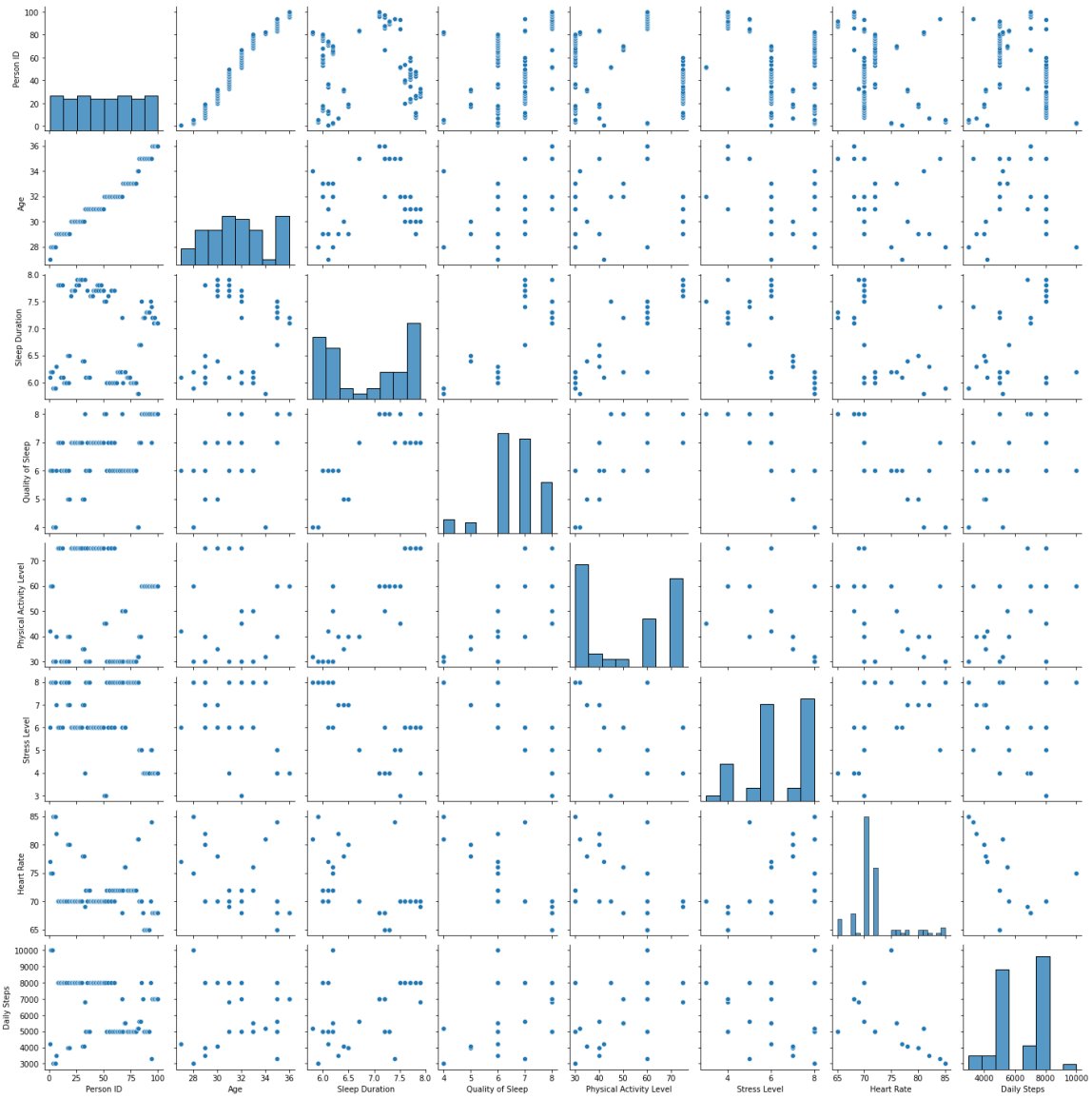
```
Index(['Person ID', 'Gender', 'Age', 'Occupation', 'Sleep Duration',
       'Quality of Sleep', 'Physical Activity Level', 'Stress Level',
       'BMI Category', 'Blood Pressure', 'Heart Rate', 'Daily Steps',
       'Sleep Disorder'],
      dtype='object')
```

In [76]:

```
sns.pairplot(b)
```

Out[76]:

```
<seaborn.axisgrid.PairGrid at 0x1b051ec8850>
```



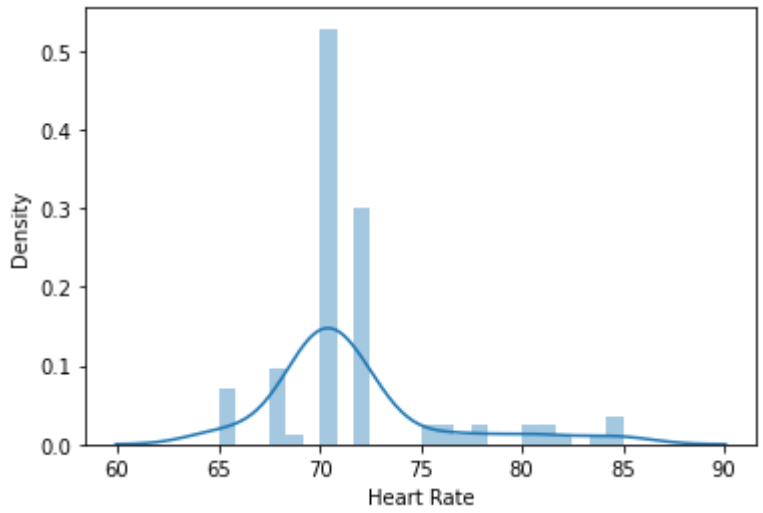
In [78]:

```
sns.distplot(b['Heart Rate'])
```

```
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[78]:

```
<AxesSubplot:xlabel='Heart Rate', ylabel='Density'>
```



In [80]:

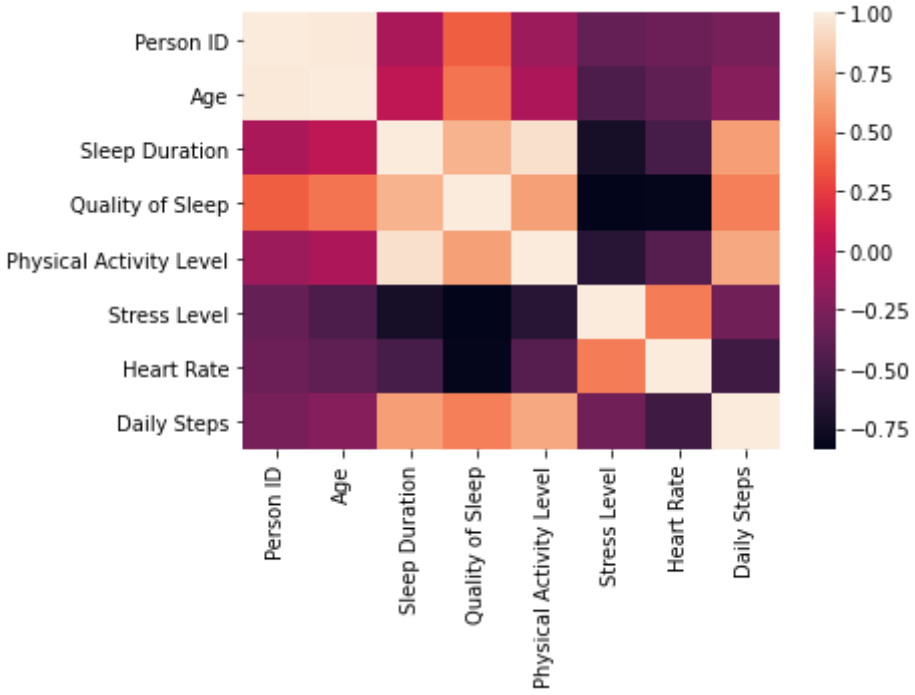
```
f=b[['Person ID', 'Gender', 'Age', 'Occupation', 'Sleep Duration',
       'Quality of Sleep', 'Physical Activity Level', 'Stress Level',
       'BMI Category', 'Blood Pressure', 'Heart Rate', 'Daily Steps',
       'Sleep Disorder']]
```

In [81]:

```
sns.heatmap(f.corr())
```

Out[81]:

<AxesSubplot:>



In [85]:

```
x=f[['Person ID','Age','Sleep Duration',
       'Quality of Sleep', 'Physical Activity Level', 'Stress Level',
       'Heart Rate', 'Daily Steps']]
y=f['Heart Rate']
```

In [86]:

```
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.5)
```

In [87]:

```
from sklearn.linear_model import LinearRegression

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

Out[87]:

```
LinearRegression()
```

In [88]:

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

```
-1.4210854715202004e-14
```

In [89]:

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

Out[89]:

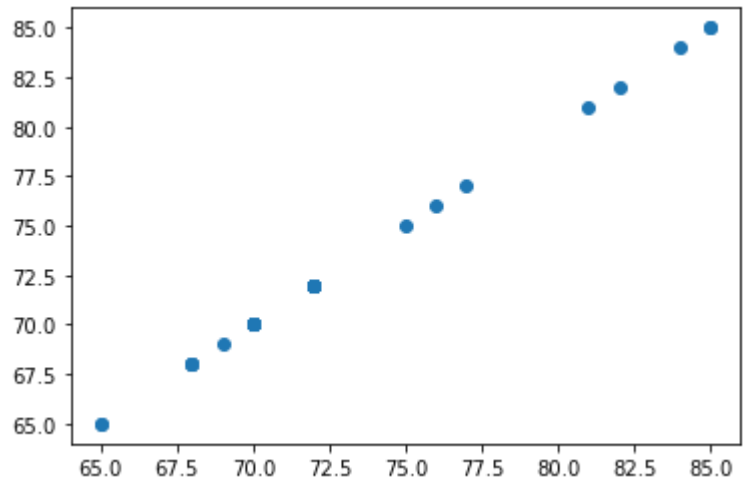
	Co-efficient
Person ID	-1.305733e-16
Age	2.810685e-15
Sleep Duration	1.190819e-15
Quality of Sleep	3.071646e-15
Physical Activity Level	1.255396e-16
Stress Level	1.093181e-15
Heart Rate	1.000000e+00
Daily Steps	-2.086921e-17

In [90]:

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

Out[90]:

```
<matplotlib.collections.PathCollection at 0x1b056b3f6a0>
```



In [91]:

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

```
1.0
```

In [92]:

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

Out[92]:

```
1.0
```

In [93]:

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

In [94]:

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

Out[94]:

```
Ridge(alpha=10)
```

In [95]:

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

Out[95]:

```
0.9990653428865417
```

In [96]:

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

Out[96]:

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

In [97]:

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

Out[97]:

```
0.5042451648341428
```

17_student_marks

In [99]:

```
a=pd.read_csv(r"C:\Users\user\Downloads\17_student_marks.csv")
a
```

Out[99]:

	Student_ID	Test_1	Test_2	Test_3	Test_4	Test_5	Test_6	Test_7	Test_8	Test_9	Test_
0	22000	78	87	91	91	88	98	94	100	100	1
1	22001	79	71	81	72	73	68	59	69	59	
2	22002	66	65	70	74	78	86	87	96	88	
3	22003	60	58	54	61	54	57	64	62	72	
4	22004	99	95	96	93	97	89	92	98	91	
5	22005	41	36	35	28	35	36	27	26	19	
6	22006	47	50	47	57	62	64	71	75	85	
7	22007	84	74	70	68	58	59	56	56	64	
8	22008	74	64	58	57	53	51	47	45	42	
9	22009	87	81	73	74	71	63	53	45	39	
10	22010	40	34	37	33	31	35	39	38	40	
11	22011	91	84	78	74	76	80	80	73	75	
12	22012	81	83	93	88	89	90	99	99	95	
13	22013	52	50	42	38	33	30	28	22	12	
14	22014	63	67	65	74	80	86	95	96	92	
15	22015	76	82	88	94	85	76	70	60	50	
16	22016	83	78	71	71	77	72	66	75	66	
17	22017	55	45	43	38	43	35	44	37	45	
18	22018	71	67	76	74	64	61	57	64	61	
19	22019	62	61	53	49	54	59	68	74	65	
20	22020	44	38	36	34	26	34	39	44	36	
21	22021	50	56	53	46	41	38	47	39	44	
22	22022	57	48	40	45	43	36	26	19	9	
23	22023	59	56	52	44	50	40	45	46	54	
24	22024	84	92	89	80	90	80	84	74	68	
25	22025	74	80	86	87	90	100	95	87	85	
26	22026	92	84	74	83	93	83	75	82	81	
27	22027	63	70	74	65	64	55	61	58	48	
28	22028	78	77	69	76	78	74	67	69	78	
29	22029	55	58	59	67	71	62	53	61	67	
30	22030	54	54	48	38	35	45	46	47	41	
31	22031	84	93	97	89	86	95	100	100	100	
32	22032	95	100	94	100	98	99	100	90	80	
33	22033	64	61	63	73	63	68	64	58	50	
34	22034	76	79	73	77	83	86	95	89	90	
35	22035	78	71	61	55	54	48	41	32	41	
36	22036	95	89	91	84	89	94	85	91	100	1

	Student_ID	Test_1	Test_2	Test_3	Test_4	Test_5	Test_6	Test_7	Test_8	Test_9	Test_
37	22037	99	89	79	87	87	81	82	74	64	
38	22038	82	83	85	86	89	80	88	95	87	
39	22039	65	56	64	62	58	51	61	68	70	
40	22040	100	93	92	86	84	76	82	74	79	
41	22041	78	72	73	79	81	73	71	77	83	
42	22042	98	100	100	93	94	92	100	100	98	
43	22043	58	62	67	77	71	63	64	73	83	
44	22044	96	92	94	100	99	95	98	92	84	
45	22045	86	87	85	84	85	91	86	82	85	
46	22046	48	55	46	40	34	29	37	34	39	
47	22047	56	52	54	47	40	35	43	44	40	
48	22048	42	44	46	53	62	59	57	53	43	
49	22049	64	54	49	59	54	55	57	59	63	
50	22050	50	44	37	29	37	46	53	57	55	
51	22051	70	60	70	62	67	67	68	67	72	
52	22052	63	73	70	63	60	67	61	59	52	
53	22053	92	100	100	100	100	100	92	87	94	1
54	22054	64	55	54	61	63	57	47	37	44	
55	22055	60	66	68	58	49	47	39	29	39	

In [100]:

```
b=a.head(100)
b
```

Out[100]:

	Student_ID	Test_1	Test_2	Test_3	Test_4	Test_5	Test_6	Test_7	Test_8	Test_9	Test_
0	22000	78	87	91	91	88	98	94	100	100	1
1	22001	79	71	81	72	73	68	59	69	59	
2	22002	66	65	70	74	78	86	87	96	88	
3	22003	60	58	54	61	54	57	64	62	72	
4	22004	99	95	96	93	97	89	92	98	91	
5	22005	41	36	35	28	35	36	27	26	19	
6	22006	47	50	47	57	62	64	71	75	85	
7	22007	84	74	70	68	58	59	56	56	64	
8	22008	74	64	58	57	53	51	47	45	42	
9	22009	87	81	73	74	71	63	53	45	39	
10	22010	40	34	37	33	31	35	39	38	40	
11	22011	91	84	78	74	76	80	80	73	75	
12	22012	81	83	93	88	89	90	99	99	95	
13	22013	52	50	42	38	33	30	28	22	12	
14	22014	63	67	65	74	80	86	95	96	92	
15	22015	76	82	88	94	85	76	70	60	50	
16	22016	83	78	71	71	77	72	66	75	66	
17	22017	55	45	43	38	43	35	44	37	45	
18	22018	71	67	76	74	64	61	57	64	61	
19	22019	62	61	53	49	54	59	68	74	65	
20	22020	44	38	36	34	26	34	39	44	36	
21	22021	50	56	53	46	41	38	47	39	44	
22	22022	57	48	40	45	43	36	26	19	9	
23	22023	59	56	52	44	50	40	45	46	54	
24	22024	84	92	89	80	90	80	84	74	68	
25	22025	74	80	86	87	90	100	95	87	85	
26	22026	92	84	74	83	93	83	75	82	81	
27	22027	63	70	74	65	64	55	61	58	48	
28	22028	78	77	69	76	78	74	67	69	78	
29	22029	55	58	59	67	71	62	53	61	67	
30	22030	54	54	48	38	35	45	46	47	41	
31	22031	84	93	97	89	86	95	100	100	100	
32	22032	95	100	94	100	98	99	100	90	80	
33	22033	64	61	63	73	63	68	64	58	50	
34	22034	76	79	73	77	83	86	95	89	90	
35	22035	78	71	61	55	54	48	41	32	41	
36	22036	95	89	91	84	89	94	85	91	100	1

	Student_ID	Test_1	Test_2	Test_3	Test_4	Test_5	Test_6	Test_7	Test_8	Test_9	Test_
37	22037	99	89	79	87	87	81	82	74	64	
38	22038	82	83	85	86	89	80	88	95	87	
39	22039	65	56	64	62	58	51	61	68	70	
40	22040	100	93	92	86	84	76	82	74	79	
41	22041	78	72	73	79	81	73	71	77	83	
42	22042	98	100	100	93	94	92	100	100	98	
43	22043	58	62	67	77	71	63	64	73	83	
44	22044	96	92	94	100	99	95	98	92	84	
45	22045	86	87	85	84	85	91	86	82	85	
46	22046	48	55	46	40	34	29	37	34	39	
47	22047	56	52	54	47	40	35	43	44	40	
48	22048	42	44	46	53	62	59	57	53	43	
49	22049	64	54	49	59	54	55	57	59	63	
50	22050	50	44	37	29	37	46	53	57	55	
51	22051	70	60	70	62	67	67	68	67	72	
52	22052	63	73	70	63	60	67	61	59	52	
53	22053	92	100	100	100	100	100	92	87	94	1
54	22054	64	55	54	61	63	57	47	37	44	
55	22055	60	66	68	58	49	47	39	29	39	

In [101]:

```
b.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 56 entries, 0 to 55
Data columns (total 13 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Student_ID  56 non-null     int64
 1   Test_1      56 non-null     int64
 2   Test_2      56 non-null     int64
 3   Test_3      56 non-null     int64
 4   Test_4      56 non-null     int64
 5   Test_5      56 non-null     int64
 6   Test_6      56 non-null     int64
 7   Test_7      56 non-null     int64
 8   Test_8      56 non-null     int64
 9   Test_9      56 non-null     int64
 10  Test_10     56 non-null     int64
 11  Test_11     56 non-null     int64
 12  Test_12     56 non-null     int64
dtypes: int64(13)
memory usage: 5.8 KB
```

In [102]:

```
b.describe()
```

Out[102]:

	Student_ID	Test_1	Test_2	Test_3	Test_4	Test_5	Test_6
count	56.000000	56.000000	56.000000	56.000000	56.000000	56.000000	56.000000
mean	22027.500000	70.750000	69.196429	68.089286	67.446429	67.303571	66.000000
std	16.309506	17.009356	17.712266	18.838333	19.807179	20.746890	21.054043
min	22000.000000	40.000000	34.000000	35.000000	28.000000	26.000000	29.000000
25%	22013.750000	57.750000	55.750000	53.000000	54.500000	53.750000	50.250000
50%	22027.500000	70.500000	68.500000	70.000000	71.500000	69.000000	65.500000
75%	22041.250000	84.000000	83.250000	85.000000	84.000000	85.250000	83.750000
max	22055.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000

In [105]:

```
b.columns
```

Out[105]:

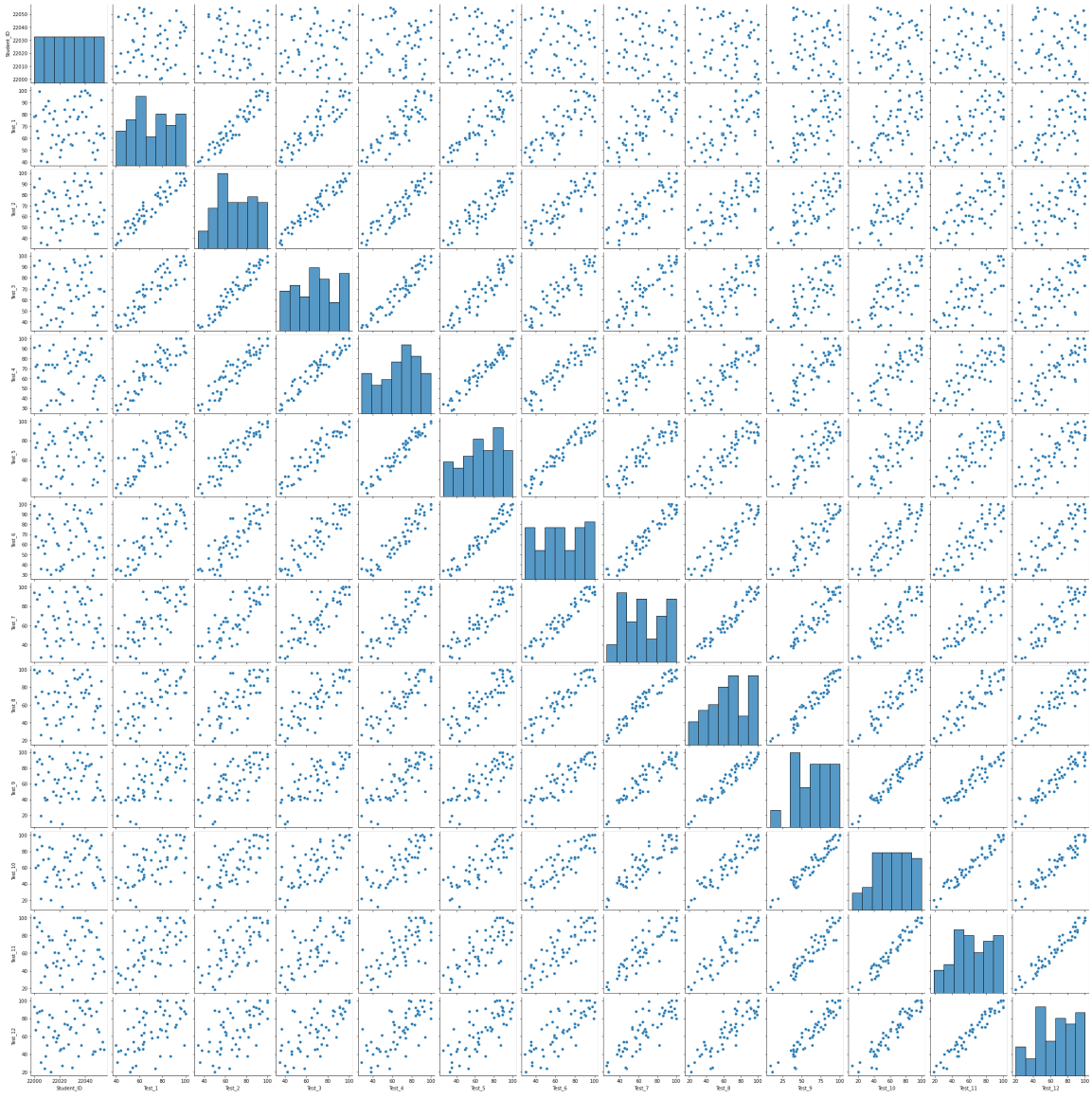
```
Index(['Student_ID', 'Test_1', 'Test_2', 'Test_3', 'Test_4', 'Test_5',
       'Test_6', 'Test_7', 'Test_8', 'Test_9', 'Test_10', 'Test_11',
       'Test_12'],
      dtype='object')
```

In [106]:

```
sns.pairplot(b)
```

Out[106]:

```
<seaborn.axisgrid.PairGrid at 0x1b055e674c0>
```



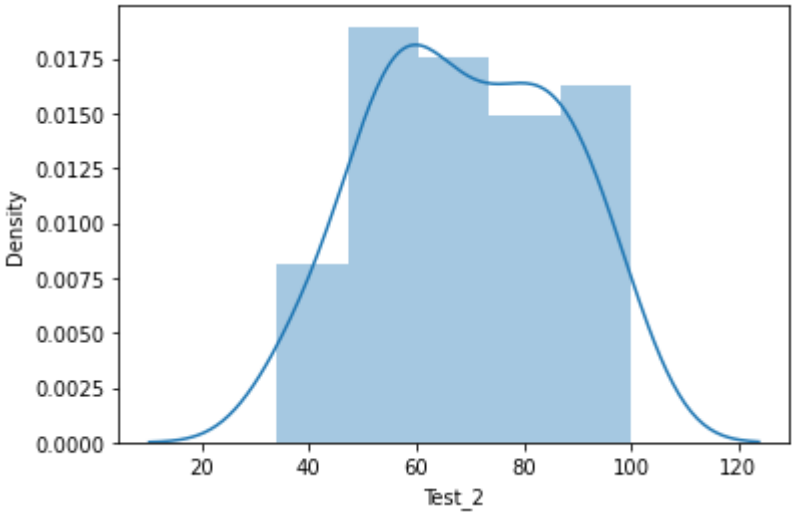
In [107]:

```
sns.distplot(b['Test_2'])
```

```
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[107]:

```
<AxesSubplot:xlabel='Test_2', ylabel='Density'>
```



In [109]:

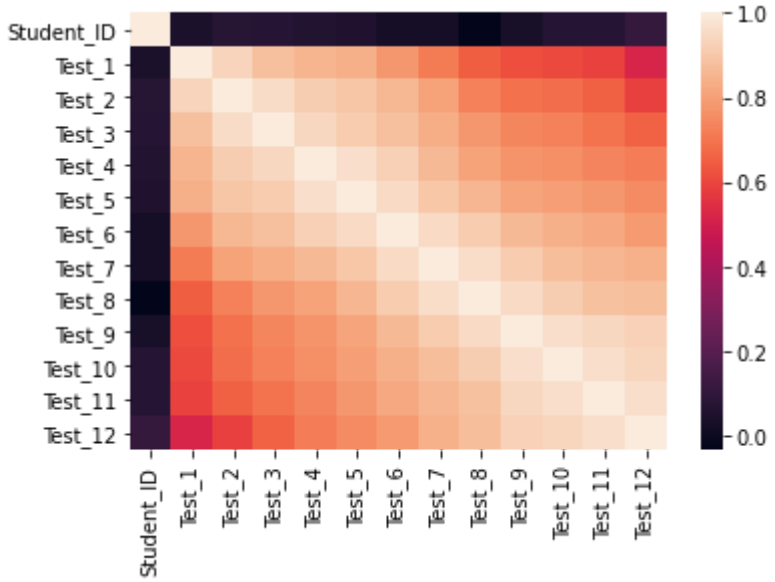
```
f=b[['Student_ID', 'Test_1', 'Test_2', 'Test_3', 'Test_4', 'Test_5',
       'Test_6', 'Test_7', 'Test_8', 'Test_9', 'Test_10', 'Test_11',
       'Test_12']]
```

In [110]:

```
sns.heatmap(f.corr())
```

Out[110]:

```
<AxesSubplot:>
```



In [113]:

```
x=f[['Student_ID', 'Test_1','Test_3', 'Test_4', 'Test_5',
       'Test_6', 'Test_7', 'Test_8', 'Test_9', 'Test_10', 'Test_11',
       'Test_12']]
y=f['Test_2']
```

In [114]:

```
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.5)
```

In [115]:

```
from sklearn.linear_model import LinearRegression

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

Out[115]:

```
LinearRegression()
```

In [116]:

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

```
-2316.9257767737035
```

In [117]:

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

Out[117]:

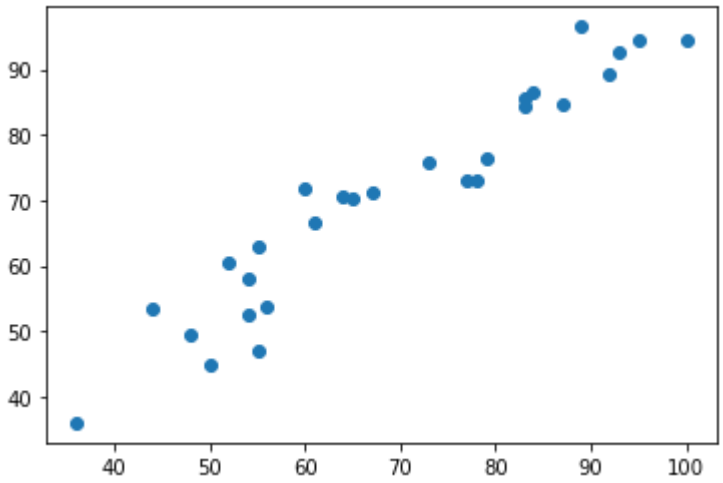
	Co-efficient
Student_ID	0.105557
Test_1	0.356388
Test_3	0.345866
Test_4	0.215177
Test_5	-0.062541
Test_6	-0.009563
Test_7	0.307048
Test_8	-0.016111
Test_9	0.013176
Test_10	-0.128335
Test_11	0.380516
Test_12	-0.502826

In [118]:

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

Out[118]:

```
<matplotlib.collections.PathCollection at 0x1b0670b9a90>
```



In [119]:

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

```
0.9078735107887352
```

In [120]:

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

Out[120]:

```
0.9877283043412819
```

In [121]:

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

In [122]:

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

Out[122]:

```
Ridge(alpha=10)
```

In [123]:

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

Out[123]:

```
0.9102198169701476
```

In [124]:

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

Out[124]:

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

In [125]:

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

Out[125]:

```
0.9437882829722327
```

18_world-data-2023

In [126]:

```
a=pd.read_csv(r"C:\Users\user\Downloads\18_world-data-2023.csv")
a
```

Out[126]:

	Country	Density\n(P/Km2)	Abbreviation	Agricultural Land(%)	Land Area(Km2)	Armed Forces size	Birth Rate	Call Co
0	Afghanistan	60	AF	58.10%	652,230	323,000	32.49	9
1	Albania	105	AL	43.10%	28,748	9,000	11.78	35
2	Algeria	18	DZ	17.40%	2,381,741	317,000	24.28	21
3	Andorra	164	AD	40.00%	468	NaN	7.20	37
4	Angola	26	AO	47.50%	1,246,700	117,000	40.73	24
...	...	...	...	...	...	...	...	
190	Venezuela	32	VE	24.50%	912,050	343,000	17.88	5
191	Vietnam	314	VN	39.30%	331,210	522,000	16.75	8
192	Yemen	56	YE	44.60%	527,968	40,000	30.45	96
193	Zambia	25	ZM	32.10%	752,618	16,000	36.19	26
194	Zimbabwe	38	ZW	41.90%	390,757	51,000	30.68	26

195 rows × 35 columns

In [127]:

```
b=a.head(100)
b
```

Out[127]:

	Country	Density\n(P/Km2)	Abbreviation	Agricultural Land(%)	Land Area(Km2)	Armed Forces size	Birth Rate	Call Cc
0	Afghanistan	60	AF	58.10%	652,230	323,000	32.49	9
1	Albania	105	AL	43.10%	28,748	9,000	11.78	35
2	Algeria	18	DZ	17.40%	2,381,741	317,000	24.28	21
3	Andorra	164	AD	40.00%	468	NaN	7.20	37
4	Angola	26	AO	47.50%	1,246,700	117,000	40.73	24
...	...	...	...	...	...	...	...	
95	Lesotho	71	LS	77.60%	30,355	2,000	26.81	26
96	Liberia	53	LR	28.00%	111,369	2,000	33.04	23
97	Libya	4	LY	8.70%	1,759,540	0	18.83	21
98	Liechtenstein	238	LI	32.20%	160	NaN	9.90	42
99	Lithuania	43	LT	47.20%	65,300	34,000	10.00	37

100 rows × 35 columns

In [128]:

```
b.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100 entries, 0 to 99
Data columns (total 35 columns):
 #   Column                                      Non-Null Count  Dtype
---  ------                                      --------------  -----
 0   Country                                     100 non-null    object
 1   Density
(P/Km2)                             100 non-null    object
 2   Abbreviation                                96 non-null     object
 3   Agricultural Land( %)                       98 non-null     object
 4   Land Area(Km2)                              100 non-null    object
 5   Armed Forces size                           92 non-null     object
 6   Birth Rate                                  98 non-null     float64
 7   Calling Code                                100 non-null    float64
 8   Capital/Major City                          99 non-null     object
 9   Co2-Emissions                               98 non-null     object
 10  CPI                                         93 non-null     object
 11  CPI Change (%)                              94 non-null     object
 12  Currency-Code                               91 non-null     object
 13  Fertility Rate                              98 non-null     float64
 14  Forested Area (%)                           98 non-null     object
 15  Gasoline Price                              93 non-null     object
 16  GDP                                         99 non-null     object
 17  Gross primary education enrollment (%)      97 non-null     object
 18  Gross tertiary education enrollment (%)     95 non-null     object
 19  Infant mortality                            97 non-null     float64
 20  Largest city                                97 non-null     object
 21  Life expectancy                             97 non-null     float64
 22  Maternal mortality ratio                    95 non-null     float64
 23  Minimum wage                                80 non-null     object
 24  Official language                           100 non-null    object
 25  Out of pocket health expenditure            97 non-null     object
 26  Physicians per thousand                     97 non-null     float64
 27  Population                                  100 non-null    object
 28  Population: Labor force participation (%)   92 non-null     object
 29  Tax revenue (%)                             87 non-null     object
 30  Total tax rate                              96 non-null     object
 31  Unemployment rate                           92 non-null     object
 32  Urban_population                            98 non-null     object
 33  Latitude                                    100 non-null    float64
 34  Longitude                                   100 non-null    float64
dtypes: float64(9), object(26)
memory usage: 27.5+ KB
```

In [129]:

```
b.describe()
```

Out[129]:

	Birth Rate	Calling Code	Fertility Rate	Infant mortality	Life expectancy	Maternal mortality ratio	Physicians per thousand	
count	98.000000	100.00000	98.000000	97.000000	97.000000	95.000000	97.000000	10(
mean	20.045000	354.15000	2.651633	21.782474	72.358763	164.305263	1.916186	2(
std	9.737986	330.66234	1.239758	20.437751	7.712838	226.488977	1.809687	2:
min	7.200000	1.00000	1.270000	1.400000	52.800000	2.000000	0.040000	-3(
25%	11.420000	84.75000	1.695000	5.000000	66.600000	12.000000	0.280000	1
50%	18.125000	253.50000	2.180000	12.800000	74.100000	58.000000	1.580000	1(
75%	27.027500	501.25000	3.322500	32.900000	78.100000	244.500000	3.190000	4
max	42.170000	1876.00000	5.920000	84.500000	84.200000	1140.000000	8.420000	6(

In [131]:

```
c=b.dropna(axis=1)
c
```

Out[131]:

	Country	Density\n(P/Km2)	Land Area(Km2)	Calling Code	Official language	Population	Latitude	Lo
0	Afghanistan	60	652,230	93.0	Pashto	38,041,754	33.939110	6
1	Albania	105	28,748	355.0	Albanian	2,854,191	41.153332	2(
2	Algeria	18	2,381,741	213.0	Arabic	43,053,054	28.033886	1
3	Andorra	164	468	376.0	Catalan	77,142	42.506285	1
4	Angola	26	1,246,700	244.0	Portuguese	31,825,295	-11.202692	1
...	...	...	...	...	...	...	...	
95	Lesotho	71	30,355	266.0	English	2,125,268	-29.609988	2(
96	Liberia	53	111,369	231.0	English	4,937,374	6.428055	-(
97	Libya	4	1,759,540	218.0	Arabic	6,777,452	26.335100	1
98	Liechtenstein	238	160	423.0	German	38,019	47.141039	(
99	Lithuania	43	65,300	370.0	Lithuanian	2,786,844	55.169438	2:

100 rows × 8 columns

In [132]:

```
c.columns
```

Out[132]:

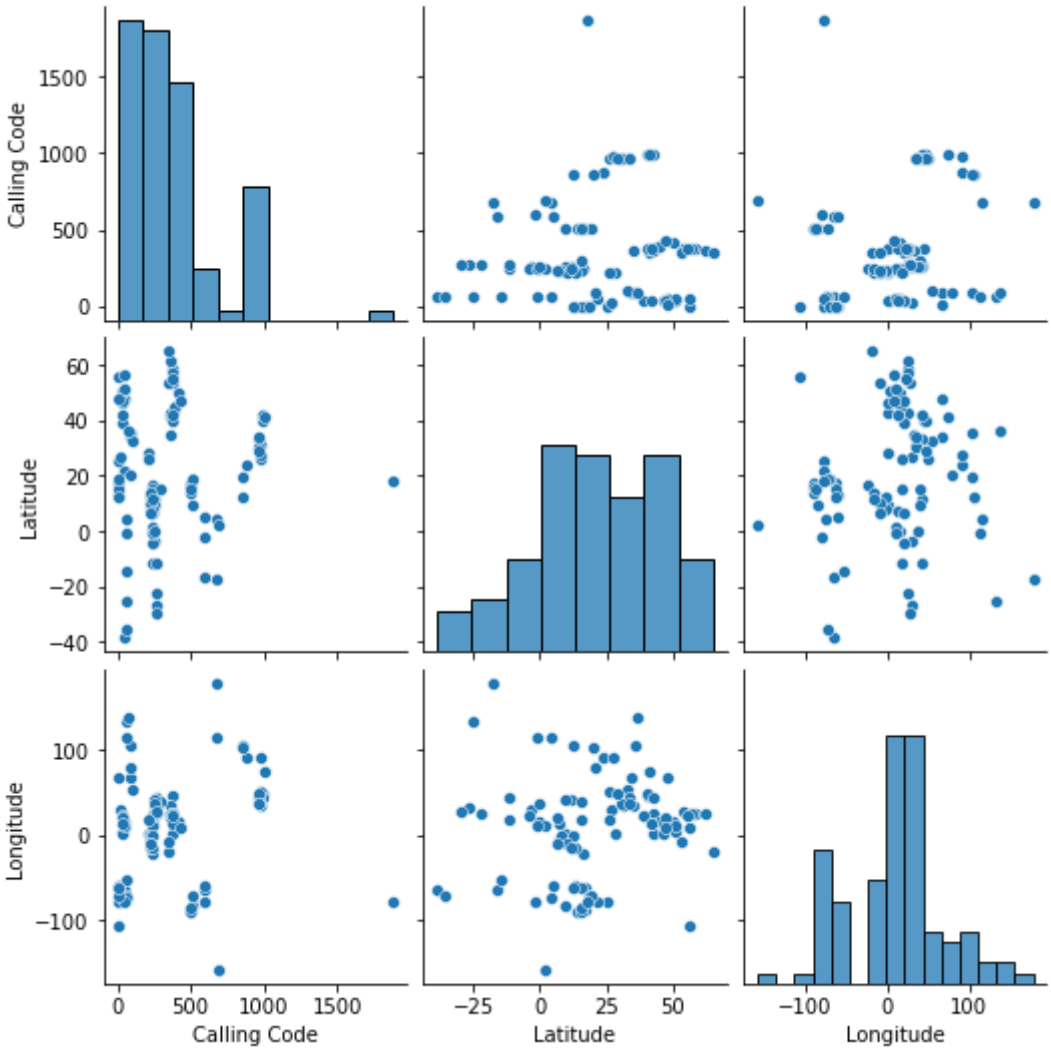
```
Index(['Country', 'Density\n(P/Km2)', 'Land Area(Km2)', 'Calling Code',
       'Official language', 'Population', 'Latitude', 'Longitude'],
      dtype='object')
```

In [133]:

```
sns.pairplot(c)
```

Out[133]:

```
<seaborn.axisgrid.PairGrid at 0x1b0670e4970>
```



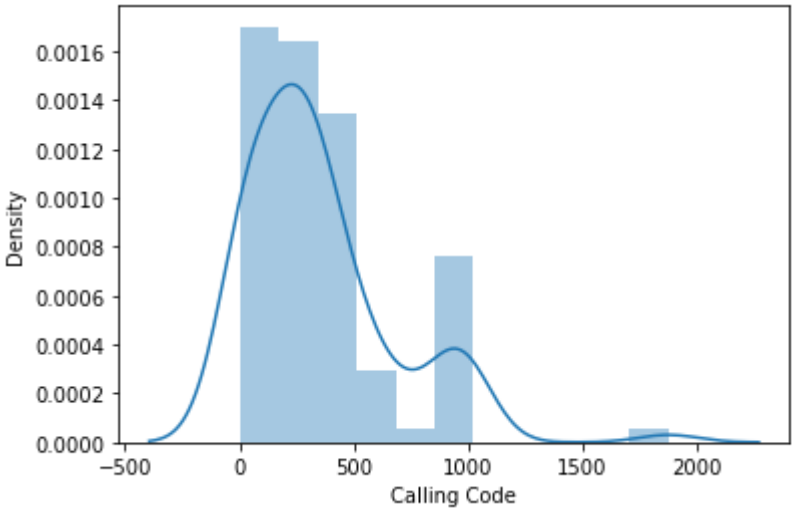
In [134]:

```
sns.distplot(c['Calling Code'])
```

```
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[134]:

```
<AxesSubplot:xlabel='Calling Code', ylabel='Density'>
```



In [135]:

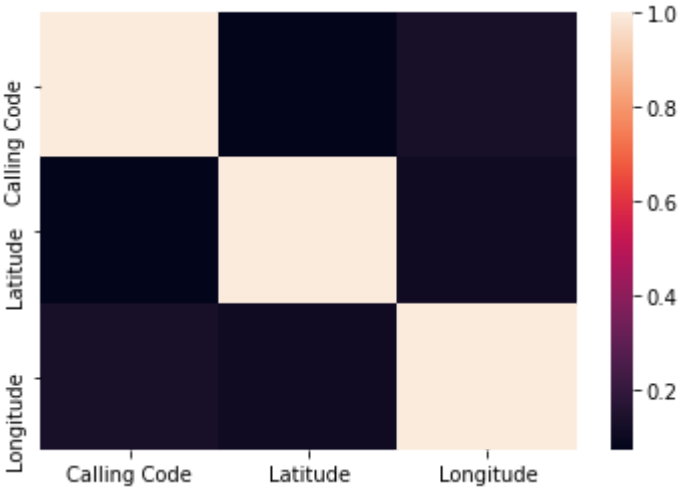
```
f=c[['Country', 'Density\n(P/Km2)', 'Land Area(Km2)', 'Calling Code',
       'Official language', 'Population', 'Latitude', 'Longitude']]
```

In [136]:

```
sns.heatmap(f.corr())
```

Out[136]:

```
<AxesSubplot:>
```



In [152]:

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

In [153]:

```
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.5)
```

In [154]:

```
from sklearn.linear_model import LinearRegression

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

Out[154]:

```
LinearRegression()
```

In [155]:

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

```
291.87882781037837
```

In [156]:

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

Out[156]:

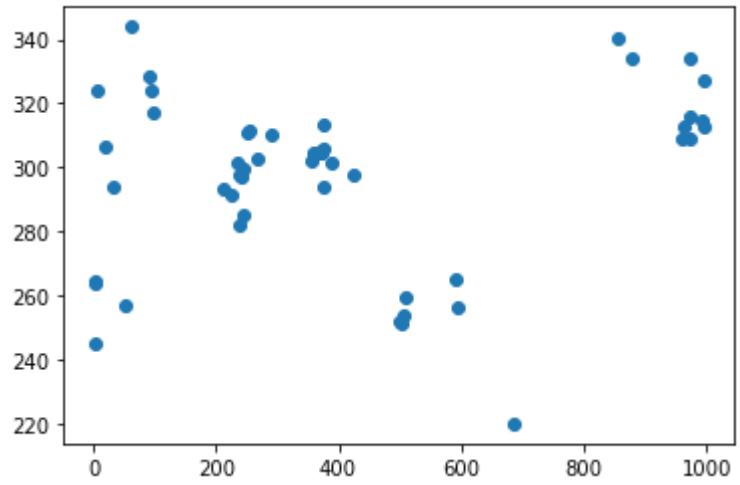
	Co-efficient
Latitude	0.027631
Longitude	0.457854

In [157]:

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

Out[157]:

```
<matplotlib.collections.PathCollection at 0x1b067871f70>
```



In [158]:

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

```
-0.10747669766799284
```

In [159]:

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

Out[159]:

```
0.006830883607632288
```

In [160]:

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

In [161]:

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

Out[161]:

```
Ridge(alpha=10)
```

In [162]:

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

Out[162]:

```
-0.10747791426514186
```

In [163]:

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

Out[163]:

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

In [164]:

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

Out[164]:

```
-0.1079837486648465
```

In [ ]:

```

```