1y54sn2uj

July 28, 2023

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

[2]:
```
df=pd.read_csv("9_bottle.csv")
df
```

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

[2]:

	Cst_Cnt	Btl_Cnt	Sta_ID	Depth_ID	\
0	1	1	054.0 056.0	19-4903CR-HY-060-0930-05400560-0000A-3	
1	1	2	054.0 056.0	19-4903CR-HY-060-0930-05400560-0008A-3	
2	1	3	054.0 056.0	19-4903CR-HY-060-0930-05400560-0010A-7	
3	1	4	054.0 056.0	19-4903CR-HY-060-0930-05400560-0019A-3	
4	1	5	054.0 056.0	19-4903CR-HY-060-0930-05400560-0020A-7	
…	…	…	…	…	
864858	34404	864859	093.4 026.4	20-1611SR-MX-310-2239-09340264-0000A-7	
864859	34404	864860	093.4 026.4	20-1611SR-MX-310-2239-09340264-0002A-3	
864860	34404	864861	093.4 026.4	20-1611SR-MX-310-2239-09340264-0005A-3	
864861	34404	864862	093.4 026.4	20-1611SR-MX-310-2239-09340264-0010A-3	
864862	34404	864863	093.4 026.4	20-1611SR-MX-310-2239-09340264-0015A-3	

	Depthm	T_degC	Salnty	O2ml_L	STheta	O2Sat	…	R_PHAEO	\
0	0	10.500	33.4400	NaN	25.64900	NaN	…	NaN	
1	8	10.460	33.4400	NaN	25.65600	NaN	…	NaN	
2	10	10.460	33.4370	NaN	25.65400	NaN	…	NaN	
3	19	10.450	33.4200	NaN	25.64300	NaN	…	NaN	
4	20	10.450	33.4210	NaN	25.64300	NaN	…	NaN	
…	…	…	…	…	…	…	…	…	
864858	0	18.744	33.4083	5.805	23.87055	108.74	…	0.18	
864859	2	18.744	33.4083	5.805	23.87072	108.74	…	0.18	
864860	5	18.692	33.4150	5.796	23.88911	108.46	…	0.18	
864861	10	18.161	33.4062	5.816	24.01426	107.74	…	0.31	
864862	15	17.533	33.3880	5.774	24.15297	105.66	…	0.61	

```
        R_PRES  R_SAMP  DIC1  DIC2  TA1  TA2  pH2  pH1  DIC Quality Comment
0            0     NaN   NaN   NaN  NaN  NaN  NaN  NaN                   NaN
1            8     NaN   NaN   NaN  NaN  NaN  NaN  NaN                   NaN
2           10     NaN   NaN   NaN  NaN  NaN  NaN  NaN                   NaN
3           19     NaN   NaN   NaN  NaN  NaN  NaN  NaN                   NaN
4           20     NaN   NaN   NaN  NaN  NaN  NaN  NaN                   NaN
…          …      …     …    …   …   …   …   …                     …
864858       0     NaN   NaN   NaN  NaN  NaN  NaN  NaN                   NaN
864859       2     4.0   NaN   NaN  NaN  NaN  NaN  NaN                   NaN
864860       5     3.0   NaN   NaN  NaN  NaN  NaN  NaN                   NaN
864861      10     2.0   NaN   NaN  NaN  NaN  NaN  NaN                   NaN
864862      15     1.0   NaN   NaN  NaN  NaN  NaN  NaN                   NaN

[864863 rows x 74 columns]
```

[3]:
```
df.head()
```

[3]:
```
   Cst_Cnt  Btl_Cnt       Sta_ID                                Depth_ID  \
0        1        1  054.0 056.0  19-4903CR-HY-060-0930-05400560-0000A-3
1        1        2  054.0 056.0  19-4903CR-HY-060-0930-05400560-0008A-3
2        1        3  054.0 056.0  19-4903CR-HY-060-0930-05400560-0010A-7
3        1        4  054.0 056.0  19-4903CR-HY-060-0930-05400560-0019A-3
4        1        5  054.0 056.0  19-4903CR-HY-060-0930-05400560-0020A-7

   Depthm  T_degC  Salnty  O2ml_L  STheta  O2Sat  …  R_PHAEO  R_PRES  \
0       0   10.50  33.440     NaN  25.649    NaN  …      NaN       0
1       8   10.46  33.440     NaN  25.656    NaN  …      NaN       8
2      10   10.46  33.437     NaN  25.654    NaN  …      NaN      10
3      19   10.45  33.420     NaN  25.643    NaN  …      NaN      19
4      20   10.45  33.421     NaN  25.643    NaN  …      NaN      20

   R_SAMP  DIC1  DIC2  TA1  TA2  pH2  pH1  DIC Quality Comment
0     NaN   NaN   NaN  NaN  NaN  NaN  NaN                   NaN
1     NaN   NaN   NaN  NaN  NaN  NaN  NaN                   NaN
2     NaN   NaN   NaN  NaN  NaN  NaN  NaN                   NaN
3     NaN   NaN   NaN  NaN  NaN  NaN  NaN                   NaN
4     NaN   NaN   NaN  NaN  NaN  NaN  NaN                   NaN

[5 rows x 74 columns]
```

1 DATA CLEANING AND DATA PREPROCESSING

[4]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 864863 entries, 0 to 864862
```

```
Data columns (total 74 columns):
 #   Column               Non-Null Count   Dtype
---  ------               --------------   -----
 0   Cst_Cnt              864863 non-null  int64
 1   Btl_Cnt              864863 non-null  int64
 2   Sta_ID               864863 non-null  object
 3   Depth_ID             864863 non-null  object
 4   Depthm               864863 non-null  int64
 5   T_degC               853900 non-null  float64
 6   Salnty               817509 non-null  float64
 7   O2ml_L               696201 non-null  float64
 8   STheta               812174 non-null  float64
 9   O2Sat                661274 non-null  float64
 10  Oxy_µmol/Kg          661268 non-null  float64
 11  BtlNum               118667 non-null  float64
 12  RecInd               864863 non-null  int64
 13  T_prec               853900 non-null  float64
 14  T_qual               23127 non-null   float64
 15  S_prec               817509 non-null  float64
 16  S_qual               74914 non-null   float64
 17  P_qual               673755 non-null  float64
 18  O_qual               184676 non-null  float64
 19  SThtaq               65823 non-null   float64
 20  O2Satq               217797 non-null  float64
 21  ChlorA               225272 non-null  float64
 22  Chlqua               639166 non-null  float64
 23  Phaeop               225271 non-null  float64
 24  Phaqua               639170 non-null  float64
 25  PO4uM                413317 non-null  float64
 26  PO4q                 451786 non-null  float64
 27  SiO3uM               354091 non-null  float64
 28  SiO3qu               510866 non-null  float64
 29  NO2uM                337576 non-null  float64
 30  NO2q                 529474 non-null  float64
 31  NO3uM                337403 non-null  float64
 32  NO3q                 529933 non-null  float64
 33  NH3uM                64962 non-null   float64
 34  NH3q                 808299 non-null  float64
 35  C14As1               14432 non-null   float64
 36  C14A1p               12760 non-null   float64
 37  C14A1q               848605 non-null  float64
 38  C14As2               14414 non-null   float64
 39  C14A2p               12742 non-null   float64
 40  C14A2q               848623 non-null  float64
 41  DarkAs               22649 non-null   float64
 42  DarkAp               20457 non-null   float64
 43  DarkAq               840440 non-null  float64
 44  MeanAs               22650 non-null   float64
```

```
 45  MeanAp                20457 non-null   float64
 46  MeanAq                840439 non-null  float64
 47  IncTim                14437 non-null   object
 48  LightP                18651 non-null   float64
 49  R_Depth               864863 non-null  float64
 50  R_TEMP                853900 non-null  float64
 51  R_POTEMP              818816 non-null  float64
 52  R_SALINITY            817509 non-null  float64
 53  R_SIGMA               812007 non-null  float64
 54  R_SVA                 812092 non-null  float64
 55  R_DYNHT               818206 non-null  float64
 56  R_O2                  696201 non-null  float64
 57  R_O2Sat               666448 non-null  float64
 58  R_SIO3                354099 non-null  float64
 59  R_PO4                 413325 non-null  float64
 60  R_NO3                 337411 non-null  float64
 61  R_NO2                 337584 non-null  float64
 62  R_NH4                 64982 non-null   float64
 63  R_CHLA                225276 non-null  float64
 64  R_PHAEO               225275 non-null  float64
 65  R_PRES                864863 non-null  int64
 66  R_SAMP                122006 non-null  float64
 67  DIC1                  1999 non-null    float64
 68  DIC2                  224 non-null     float64
 69  TA1                   2084 non-null    float64
 70  TA2                   234 non-null     float64
 71  pH2                   10 non-null      float64
 72  pH1                   84 non-null      float64
 73  DIC Quality Comment   55 non-null      object
dtypes: float64(65), int64(5), object(4)
memory usage: 488.3+ MB
```

[5]:
```
df.describe()
```

[5]:

	Cst_Cnt	Btl_Cnt	Depthm	T_degC	\
count	864863.000000	864863.000000	864863.000000	853900.000000	
mean	17138.790958	432432.000000	226.831951	10.799677	
std	10240.949817	249664.587267	316.050259	4.243825	
min	1.000000	1.000000	0.000000	1.440000	
25%	8269.000000	216216.500000	46.000000	7.680000	
50%	16848.000000	432432.000000	125.000000	10.060000	
75%	26557.000000	648647.500000	300.000000	13.880000	
max	34404.000000	864863.000000	5351.000000	31.140000	

	Salnty	O2ml_L	STheta	O2Sat	\
count	817509.000000	696201.000000	812174.000000	661274.000000	
mean	33.840350	3.392468	25.819394	57.103779	

```
std          0.461843       2.073256       1.167787      37.094137
min         28.431000      -0.010000      20.934000      -0.100000
25%         33.488000       1.360000      24.965000      21.100000
50%         33.863000       3.440000      25.996000      54.400000
75%         34.196900       5.500000      26.646000      97.600000
max         37.034000      11.130000     250.784000     214.100000

          Oxy_µmol/Kg         BtlNum  …         R_CHLA        R_PHAEO  \
count  661268.000000  118667.000000  …  225276.000000  225275.000000
mean      148.808694      10.497426  …       0.450225       0.198599
std        90.187533       6.189688  …       1.208566       0.376539
min        -0.434900       0.000000  …      -0.010000      -3.890000
25%        60.915470       5.000000  …       0.050000       0.050000
50%       151.064150      10.000000  …       0.160000       0.110000
75%       240.379600      16.000000  …       0.390000       0.230000
max       485.701800      25.000000  …      66.110000      65.300000

              R_PRES         R_SAMP         DIC1         DIC2          TA1  \
count  864863.000000  122006.000000  1999.000000   224.000000  2084.000000
mean      228.395694     162.071521  2153.239714  2168.148330  2256.055845
std       319.456731      85.722796   112.995202   154.852332    34.844435
min         0.000000       0.000000  1948.850000  1969.440000  2181.570000
25%        46.000000     200.000000  2028.330000  2008.977500  2230.322500
50%       126.000000     206.000000  2170.640000  2265.885000  2244.325000
75%       302.000000     214.000000  2253.810000  2315.525000  2278.505000
max      5458.000000     424.000000  2367.800000  2364.420000  2434.900000

               TA2        pH2        pH1
count   234.000000  10.000000  84.000000
mean   2278.858803   7.948570   7.910983
std      58.496495   0.021216   0.077666
min    2198.150000   7.923100   7.618300
25%    2229.062500   7.931475   7.898675
50%    2247.505000   7.946650   7.928850
75%    2316.452500   7.963300   7.955100
max    2437.000000   7.988300   8.047700

[8 rows x 70 columns]
```

[6]:
```
df.columns
```

[6]:
```
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')
```

[7]:
```
df1=df.dropna(axis=1)
df1
```

[7]:

	Cst_Cnt	Btl_Cnt	Sta_ID	Depth_ID	\
0	1	1	054.0 056.0	19-4903CR-HY-060-0930-05400560-0000A-3	
1	1	2	054.0 056.0	19-4903CR-HY-060-0930-05400560-0008A-3	
2	1	3	054.0 056.0	19-4903CR-HY-060-0930-05400560-0010A-7	
3	1	4	054.0 056.0	19-4903CR-HY-060-0930-05400560-0019A-3	
4	1	5	054.0 056.0	19-4903CR-HY-060-0930-05400560-0020A-7	
…	…	…	…	…	
864858	34404	864859	093.4 026.4	20-1611SR-MX-310-2239-09340264-0000A-7	
864859	34404	864860	093.4 026.4	20-1611SR-MX-310-2239-09340264-0002A-3	
864860	34404	864861	093.4 026.4	20-1611SR-MX-310-2239-09340264-0005A-3	
864861	34404	864862	093.4 026.4	20-1611SR-MX-310-2239-09340264-0010A-3	
864862	34404	864863	093.4 026.4	20-1611SR-MX-310-2239-09340264-0015A-3	

	Depthm	RecInd	R_Depth	R_PRES
0	0	3	0.0	0
1	8	3	8.0	8
2	10	7	10.0	10
3	19	3	19.0	19
4	20	7	20.0	20
…	…	…	…	
864858	0	7	0.0	0
864859	2	3	2.0	2
864860	5	3	5.0	5
864861	10	3	10.0	10
864862	15	3	15.0	15

```
[864863 rows x 8 columns]
```

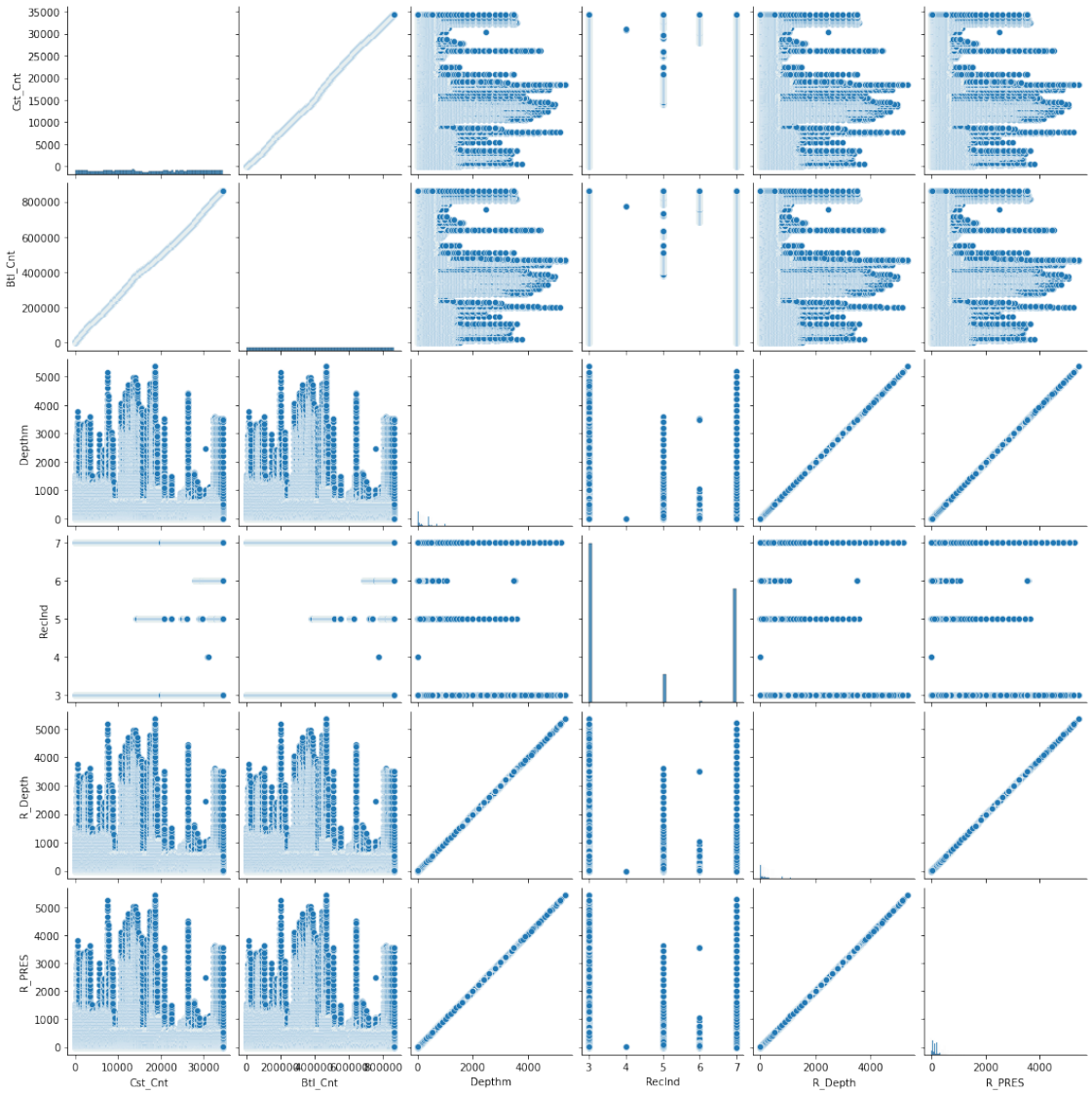
[8]:
```
df1.columns
```

[8]:
```
Index(['Cst_Cnt', 'Btl_Cnt', 'Sta_ID', 'Depth_ID', 'Depthm', 'RecInd',
       'R_Depth', 'R_PRES'],
      dtype='object')
```

2 EDA AND VISUALIZATION

[9]:
```
sns.pairplot(df1)
```

[9]:
```
<seaborn.axisgrid.PairGrid at 0x217cce72280>
```

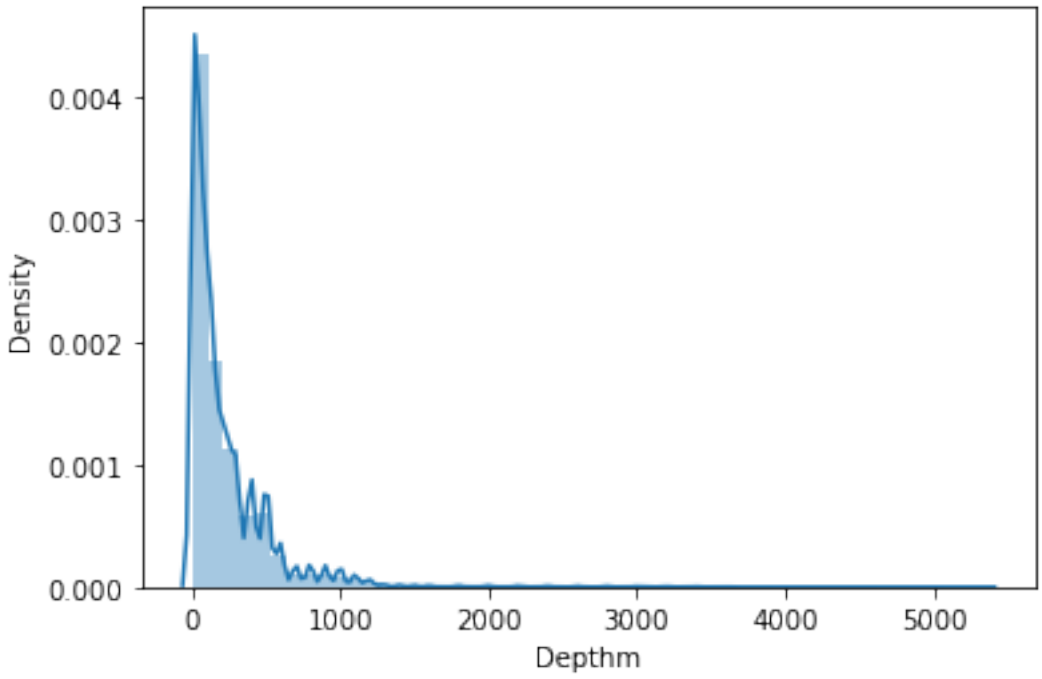


[10]:
```
sns.distplot(df1['Depthm'])
```

```
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 function for
histograms).
```

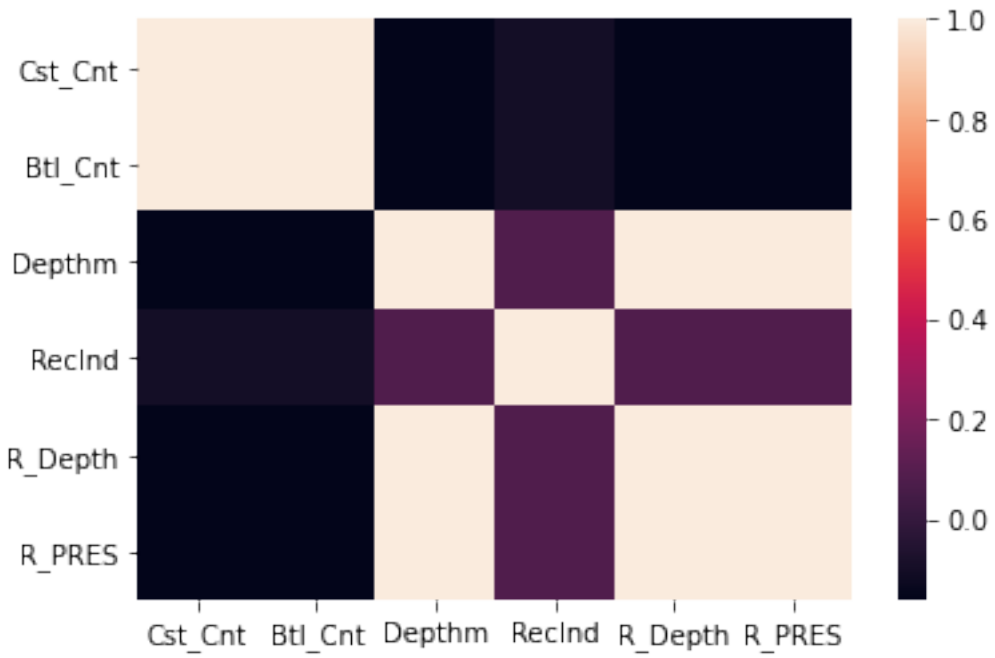
```
  warnings.warn(msg, FutureWarning)
```

[10]: `<AxesSubplot:xlabel='Depthm', ylabel='Density'>`



[11]:
```
sns.heatmap(df1.corr())
```

[11]: `<AxesSubplot:>`



3 TO TRAIN THE MODEL AND MODEL BULDING

[12]:
```
x=df[['Cst_Cnt', 'Btl_Cnt', 'Depthm', 'RecInd','R_Depth']]
y=df['R_PRES']
```

[13]:
```
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)
```

[14]:
```
from sklearn.linear_model import LinearRegression
lr=LinearRegression()
lr.fit(x_train,y_train)
```

[14]:
```
LinearRegression()
```

[15]:
```
lr.intercept_
```

[15]:
```
-1.0516771507653857
```

[16]:
```
coeff=pd.DataFrame(lr.coef_,x.columns,columns=['Co-efficient'])
coeff
```

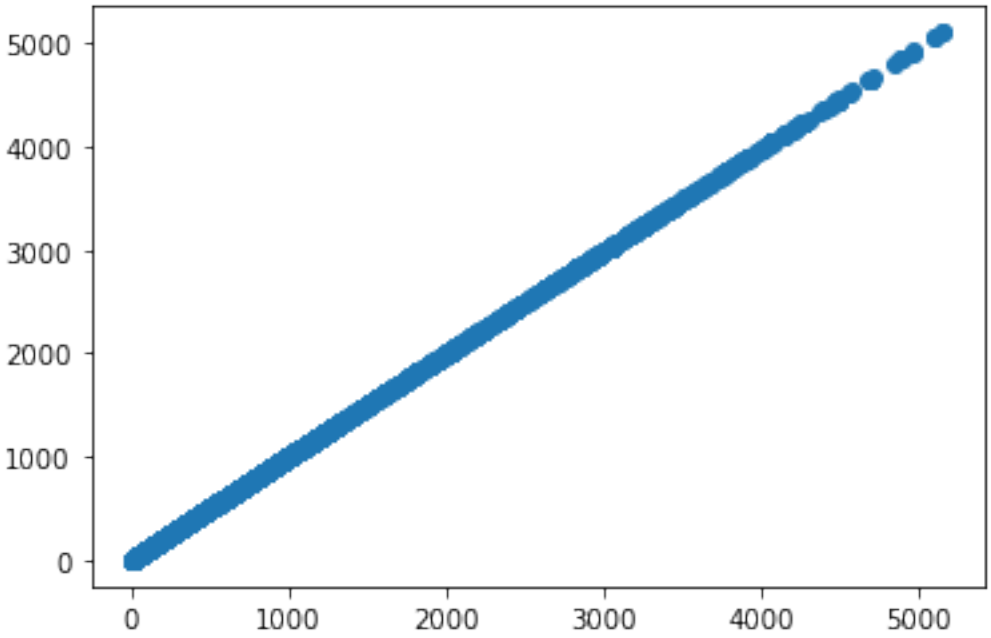
[16]:

	Co-efficient
Cst_Cnt	-0.000167
Btl_Cnt	0.000007

```
Depthm      -0.758716
RecInd      -0.018884
R_Depth      1.769554
```

[17]:
```
prediction =lr.predict(x_test)
plt.scatter(y_test,prediction)
```

[17]: <matplotlib.collections.PathCollection at 0x217861a2190>



4 ACCURACY

[18]:
```
lr.score(x_test,y_test)
```

[18]: 0.9999882492737853

[19]:
```
lr.score(x_train,y_train)
```

[19]: 0.9999878713294011

[20]:
```
from sklearn.linear_model import Ridge,Lasso
```

[21]:
```
rr=Ridge(alpha=10)
rr.fit(x_train,y_train)
```

[21]: Ridge(alpha=10)

[22]:
```
rr.score(x_test,y_test)
```

[22]: 0.999988249433992

[23]:
```
rr.score(x_train,y_train)
```

[23]: 0.9999878713015609

[24]:
```
la=Lasso(alpha=10)
la.fit(x_train,y_train)
```

[24]: Lasso(alpha=10)

[25]:
```
la.score(x_train,y_train)
```

[25]: 0.999987797565805

[26]:
```
la.score(x_test,y_test)
```

[26]: 0.9999881933268963