To prediction Crop yield

Importing Important Libraries

```
import pandas as pd
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score, mean_squared_error as mse
import warnings
warnings.filterwarnings("ignore")
import matplotlib.pyplot as plt
import numpy as np

z = pd.read_csv(r"C:\Users\skj_h\OneDrive\Desktop\crop yield
prediction dataset\train.csv")
z
```

```
    id  Year   State Crop_Type  Rainfall Soil_Type  Irrigation_Area  \
0    1  2019  Punjab     Wheat     578.6     Loamy           3515.2
1    2  2018  Punjab     Wheat     598.3     Loamy           3499.3
2    3  2017  Punjab     Wheat     493.0     Loamy           3467.7
3    4  2016  Punjab     Wheat     426.7     Loamy           3474.6
4    5  2015  Punjab     Wheat     546.9     Loamy           3474.7
5    6  2014  Punjab     Wheat     384.9     Loamy           3474.7
6    7  2013  Punjab     Wheat     619.7     Loamy           3488.1
7    8  2011  Punjab     Wheat     218.9     Loamy           3466.9
8    9  2010  Punjab     Wheat     472.1     Loamy           3474.8
9   10  2009  Punjab     Wheat     384.9     Loamy           3474.8
10  11  2008  Punjab     Wheat     529.2     Loamy           3437.9
11  12  2007  Punjab     Wheat     438.0     Loamy           3406.9
12  13  2006  Punjab     Wheat     418.3     Loamy           3404.8
13  14  2005  Punjab     Wheat     565.9     Loamy           3410.5
14  15  2004  Punjab     Wheat     375.2     Loamy           3381.7
15  16  2003  Punjab     Wheat     459.5     Loamy           3311.6
16  17  2002  Punjab     Wheat     314.5     Loamy           3353.5
17  18  2001  Punjab     Wheat     662.8     Loamy           3333.6
18  19  2000  Punjab     Wheat     391.9     Loamy           3284.3
19  20  2021  Punjab      Rice     556.9  alluvial           3229.5
20  21  2020  Punjab      Rice     602.6  alluvial           3118.8
21  22  2016  Punjab      Rice     426.7  alluvial           2961.4
22  23  2015  Punjab      Rice     546.9  alluvial           2838.3
23  24  2014  Punjab      Rice     384.9  alluvial           2838.3
24  25  2013  Punjab      Rice     619.7  alluvial           2837.6
25  26  2011  Punjab      Rice     218.9  alluvial           2814.2
26  27  2010  Punjab      Rice     472.1  alluvial           2721.8
27  28  2009  Punjab      Rice     384.9  alluvial           2721.8
28  29  2008  Punjab      Rice     529.2  alluvial           2592.2
29  30  2007  Punjab      Rice     438.0  alluvial           2602.4
30  31  2006  Punjab      Rice     418.3  alluvial           2639.9
```

```
31  32  2005  Punjab      Rice     565.9  alluvial          2632.3
32  33  2004  Punjab      Rice     375.2  alluvial          2599.6
33  34  2003  Punjab      Rice     459.5  alluvial          2515.7
34  35  2002  Punjab      Rice     314.5  alluvial          2471.0
35  36  2001  Punjab      Rice     662.8  alluvial          2590.3
36  37  2000  Punjab      Rice     391.9  alluvial          2584.7
37  38  2021  Punjab     Bajra     556.9     Loamy             3.9
38  39  2020  Punjab     Bajra     602.6     Loamy             2.0
39  40  2019  Punjab     Bajra     578.6     Loamy             1.9
40  41  2018  Punjab     Bajra     598.3     Loamy             2.8
41  42  2017  Punjab     Bajra     493.0     Loamy             3.1
42  43  2016  Punjab     Bajra     426.7     Loamy             1.9
43  44  2015  Punjab     Bajra     546.9     Loamy             1.2
44  45  2010  Punjab     Bajra     472.1     Loamy             4.9
45  46  2009  Punjab     Bajra     384.9     Loamy             4.9
46  47  2008  Punjab     Bajra     529.2     Loamy             3.5
47  48  2007  Punjab     Bajra     438.0     Loamy             4.2
48  49  2006  Punjab     Bajra     418.3     Loamy             5.2
49  50  2005  Punjab     Bajra     565.9     Loamy             5.6
50  51  2004  Punjab     Bajra     375.2     Loamy             7.2
51  52  2003  Punjab     Bajra     459.5     Loamy             6.1
52  53  2002  Punjab     Bajra     314.5     Loamy             7.6
53  54  2001  Punjab     Bajra     662.8     Loamy             5.4
54  55  2000  Punjab     Bajra     391.9     Loamy             4.6

    Crop_Yield (kg/ha)
0                 5188
1                 5077
2                 5046
3                 4583
4                 4304
5                 5017
6                 4724
7                 4693
8                 4307
9                 4462
10                4507
11                4210
12                4179
13                4221
14                4207
15                4200
16                4532
17                4563
18                4696
19                4443
20                4034
21                3974
22                3838
```

```
23                   3952
24                   3998
25                   3828
26                   4010
27                   4022
28                   4019
29                   3868
30                   3858
31                   3943
32                   3694
33                   3510
34                   3545
35                   3506
36                   3347
37                     40
38                    635
39                    583
40                    597
41                    580
42                      0
43                      0
44                   1495
45                   1055
46                    950
47                    977
48                   1045
49                    978
50                    993
51                    810
52                    929
53                    893
54                    703
```

```
z.isnull().sum()
```

```
id                    0
Year                  0
State                 0
Crop_Type             0
Rainfall              0
Soil_Type             0
Irrigation_Area       0
Crop_Yield (kg/ha)    0
dtype: int64
```

```
z.shape
```

```
(55, 8)
```

```
z.size
```

```
440
```

```
z.describe()
```

```
               id         Year    Rainfall  Irrigation_Area  Crop_Yield
(kg/ha)
count  55.00000    55.000000   55.000000        55.000000
55.000000
mean   28.00000  2009.527273  473.881818      2082.207273
3079.418182
std    16.02082     6.394021  106.836760      1495.190498
1706.608372
min     1.00000  2000.000000  218.900000         1.200000
0.000000
25%    14.50000  2004.000000  391.900000         5.500000
985.500000
50%    28.00000  2009.000000  459.500000      2721.800000
3943.000000
75%    41.50000  2015.000000  561.400000      3393.250000
4305.500000
max    55.00000  2021.000000  662.800000      3515.200000
5188.000000
```

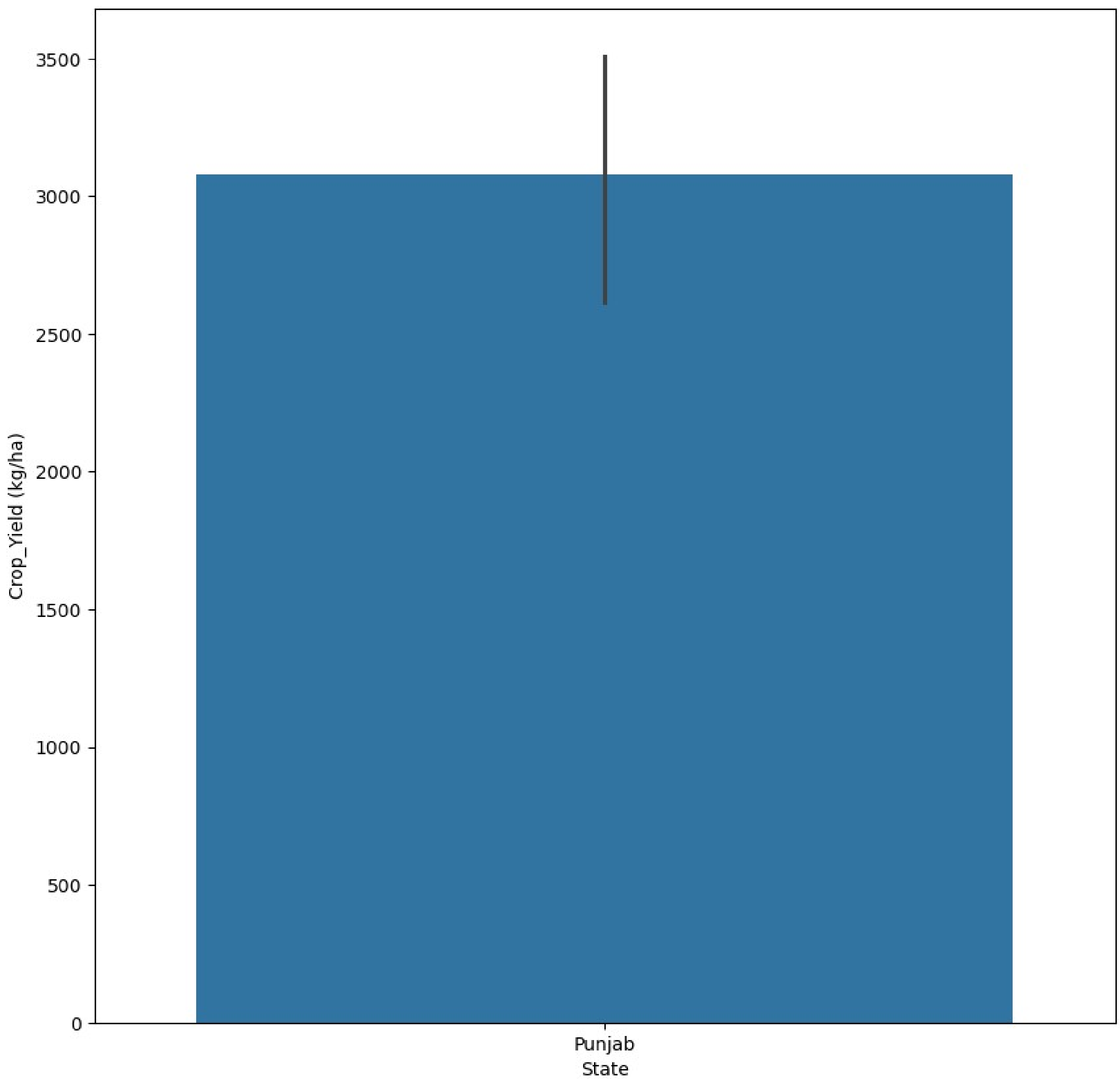
```
z.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 55 entries, 0 to 54
Data columns (total 8 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   id                  55 non-null     int64  
 1   Year                55 non-null     int64  
 2   State               55 non-null     object 
 3   Crop_Type           55 non-null     object 
 4   Rainfall            55 non-null     float64
 5   Soil_Type           55 non-null     object 
 6   Irrigation_Area     55 non-null     float64
 7   Crop_Yield (kg/ha)  55 non-null     int64  
dtypes: float64(2), int64(3), object(3)
memory usage: 3.6+ KB
```

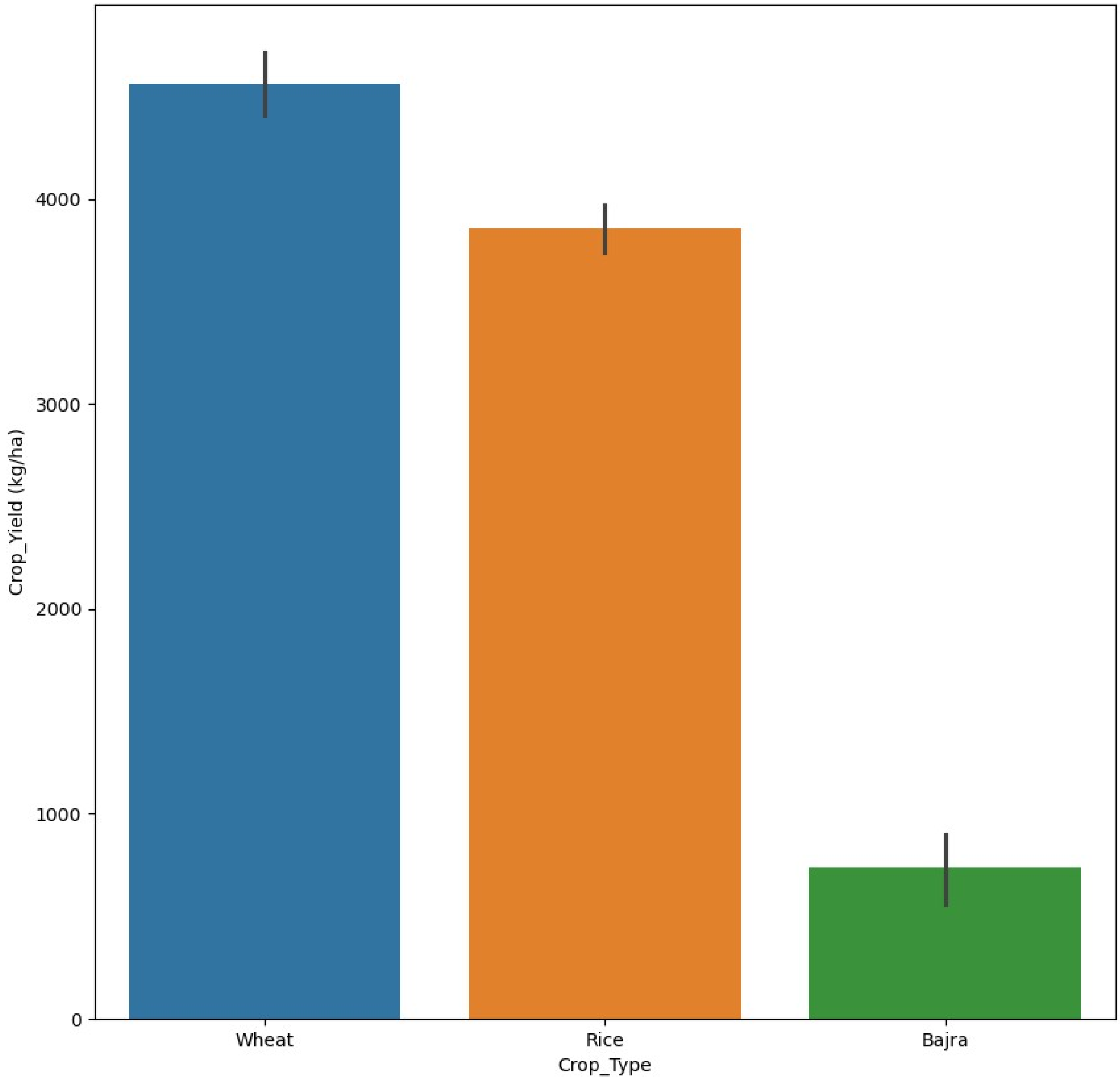
```
z["id"] = z["id"].astype(str)
```

Data Analysis

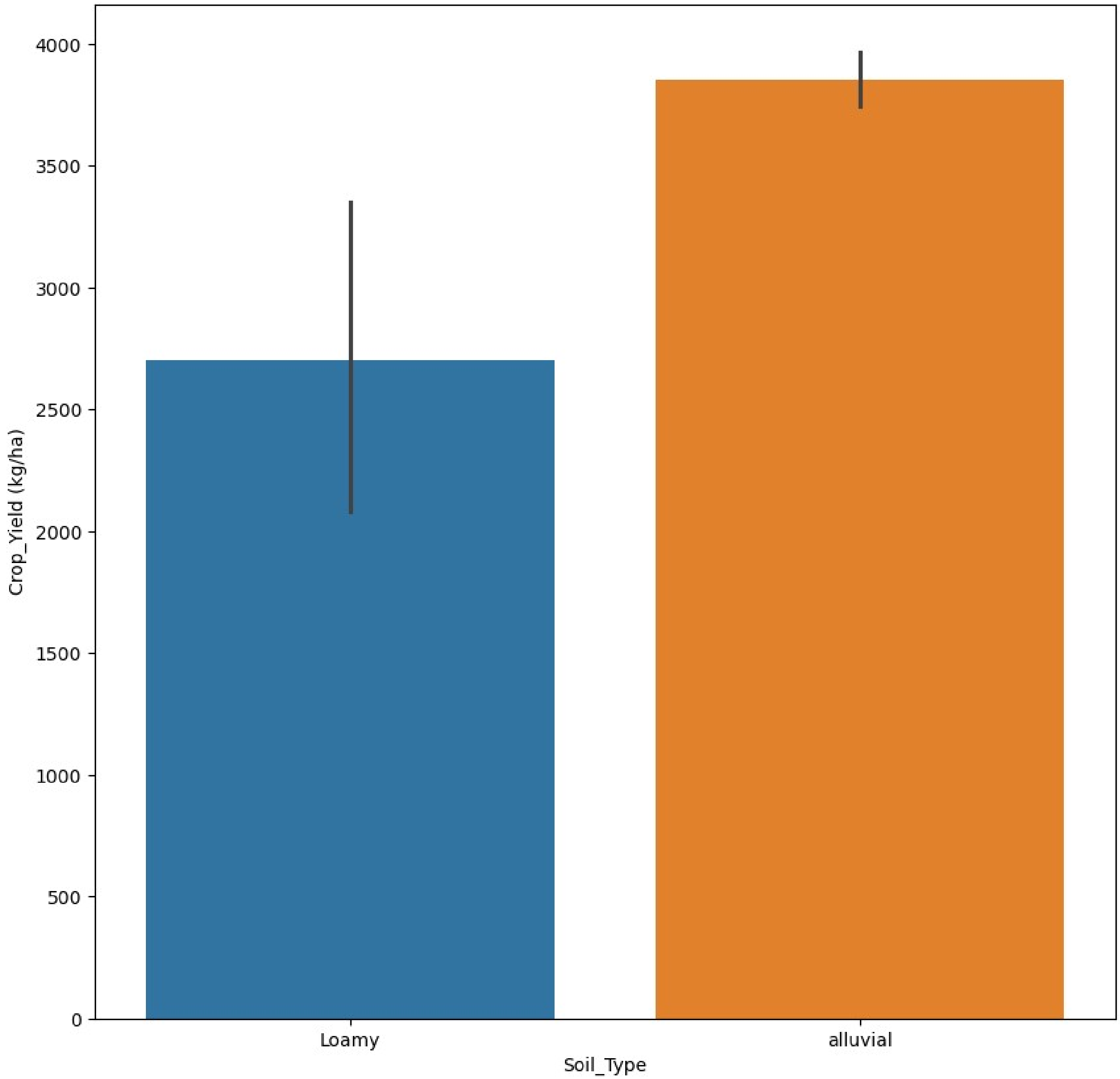
```
for i in z.columns:
    if(z[i].dtype == "object") and ( i != "id"):
        plt.figure(figsize = (10, 10))
        sns.barplot(x = z[i], y = z["Crop_Yield (kg/ha)"], data = z, 
hue = z[i])
```



3500
3000
2500
2000
1500
1000
500
0
Crop_Yield (kg/ha)
Punjab
State

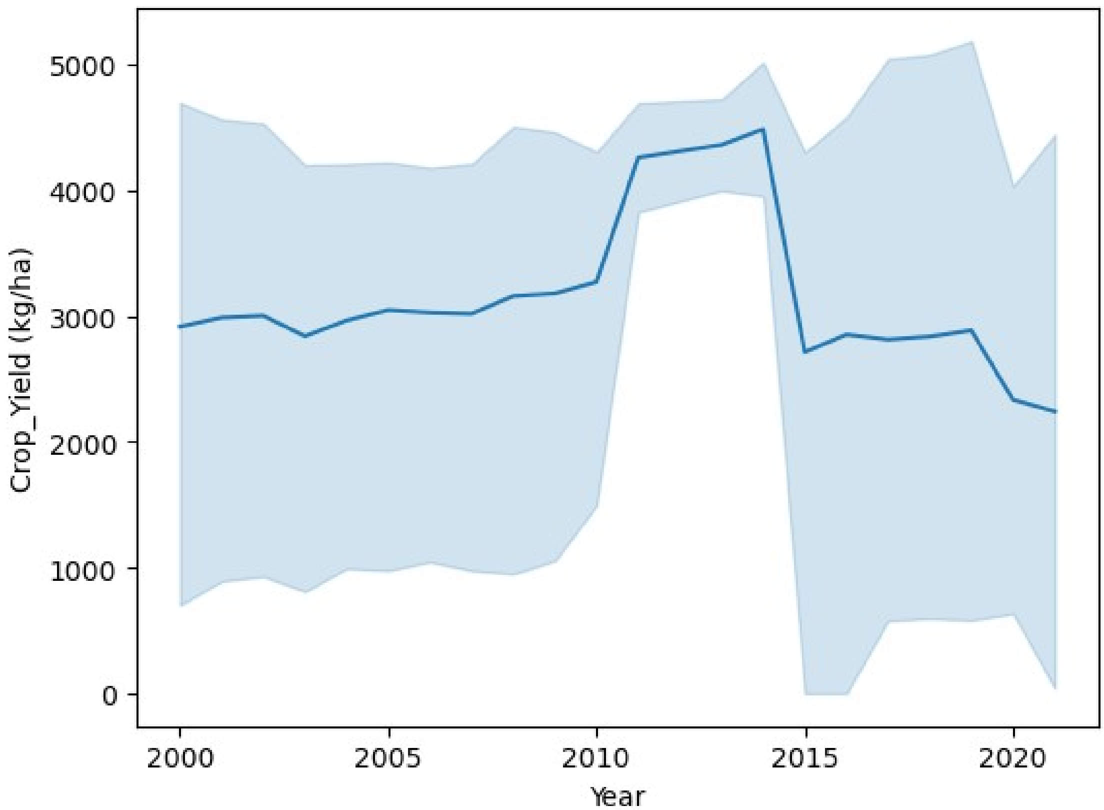


4000
3000
2000
1000
0
Crop_Yield (kg/ha)
Wheat
Rice
Bajra
Crop_Type

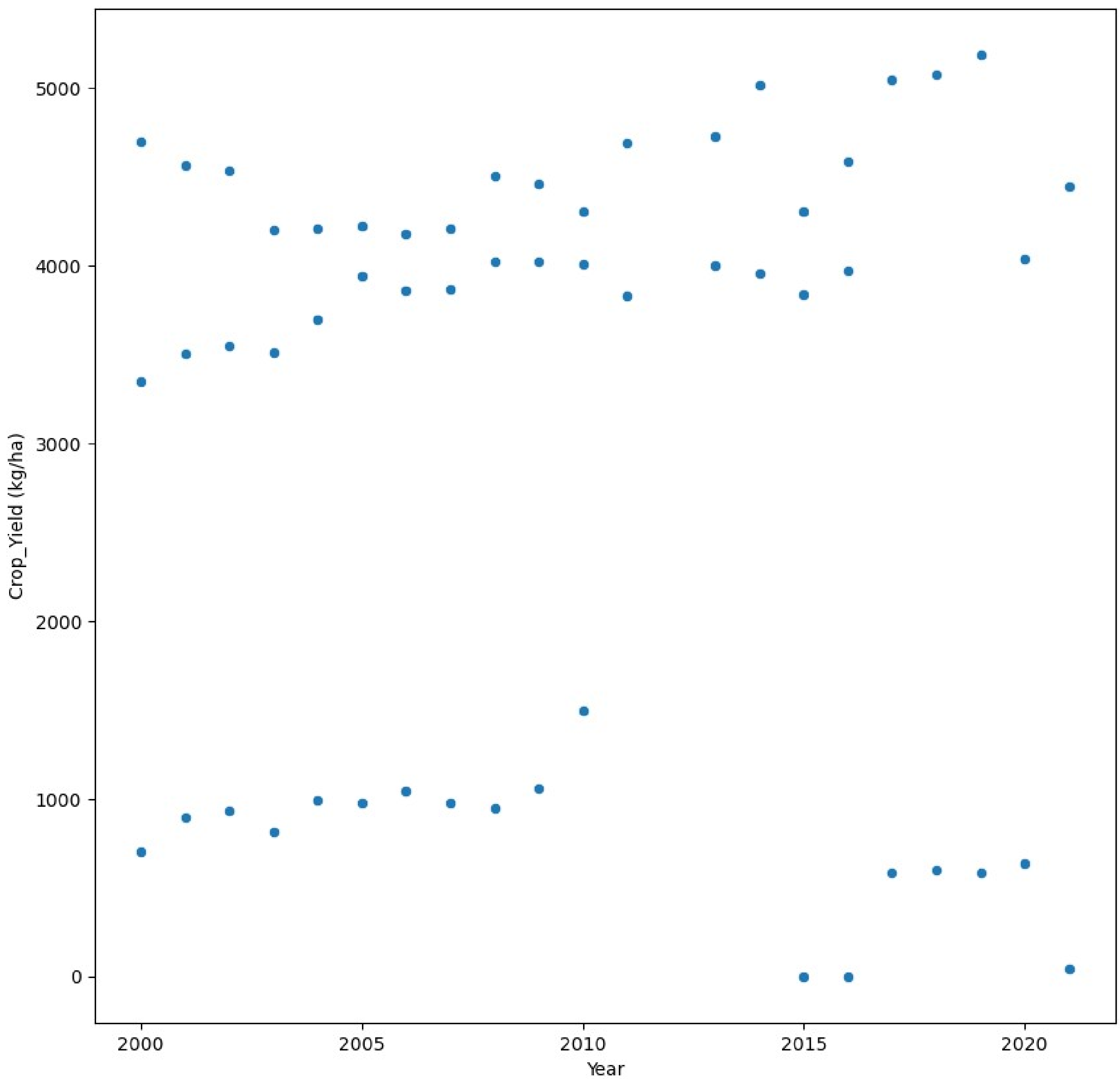



```
sns.lineplot(x = z["Year"], y = z["Crop_Yield (kg/ha)"], data = z)
```

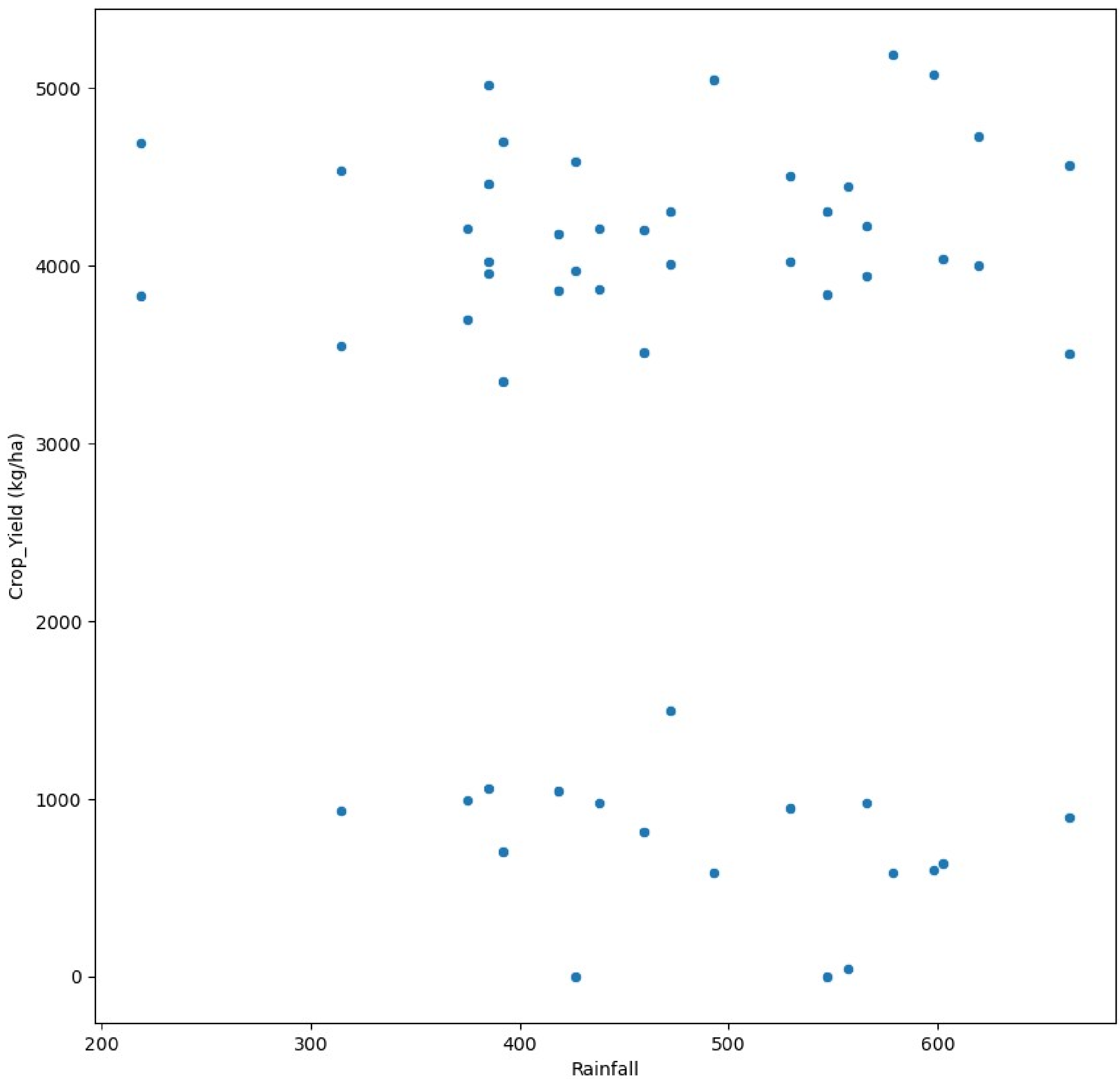
```
<Axes: xlabel='Year', ylabel='Crop_Yield (kg/ha)'>
```

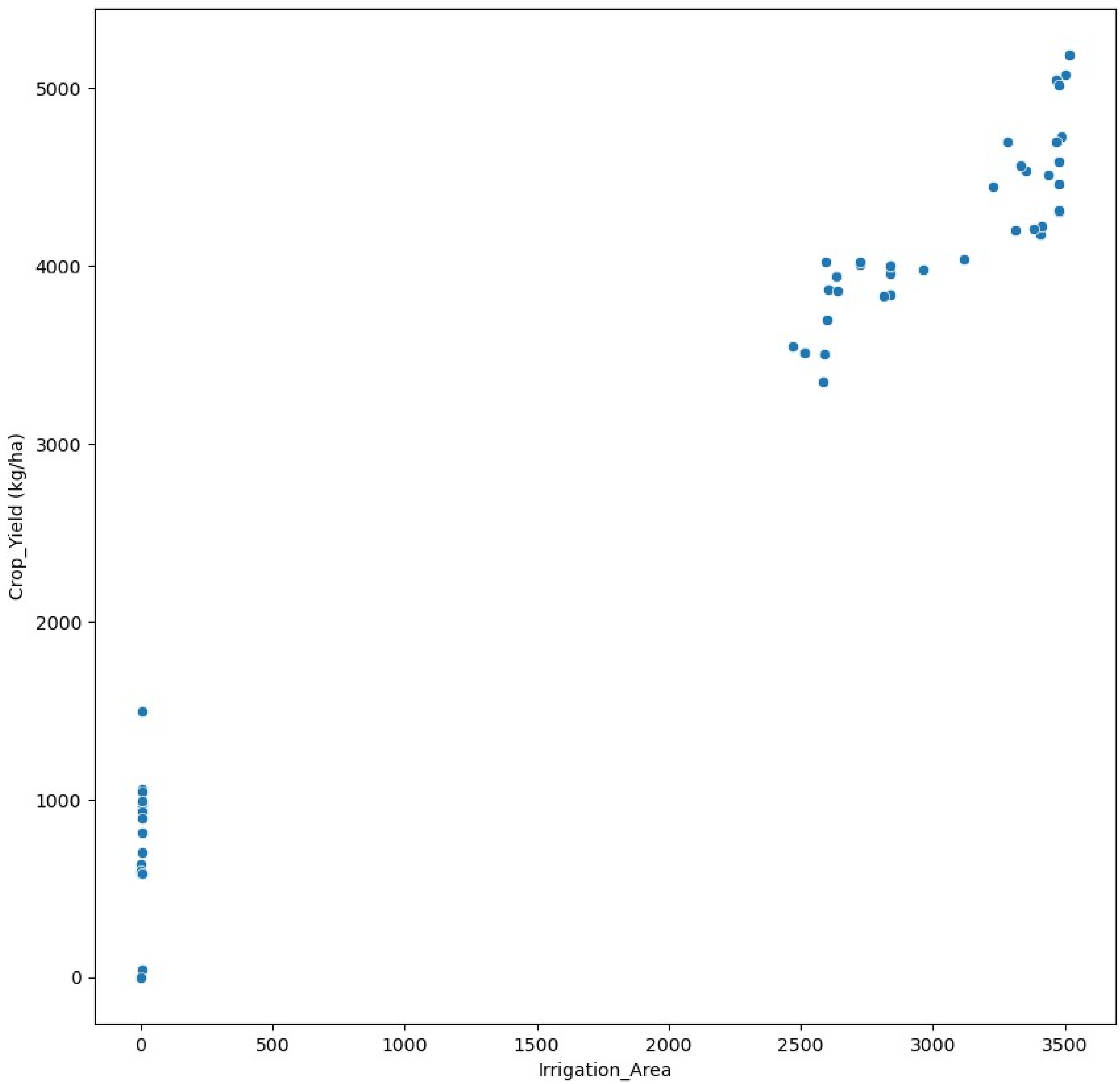


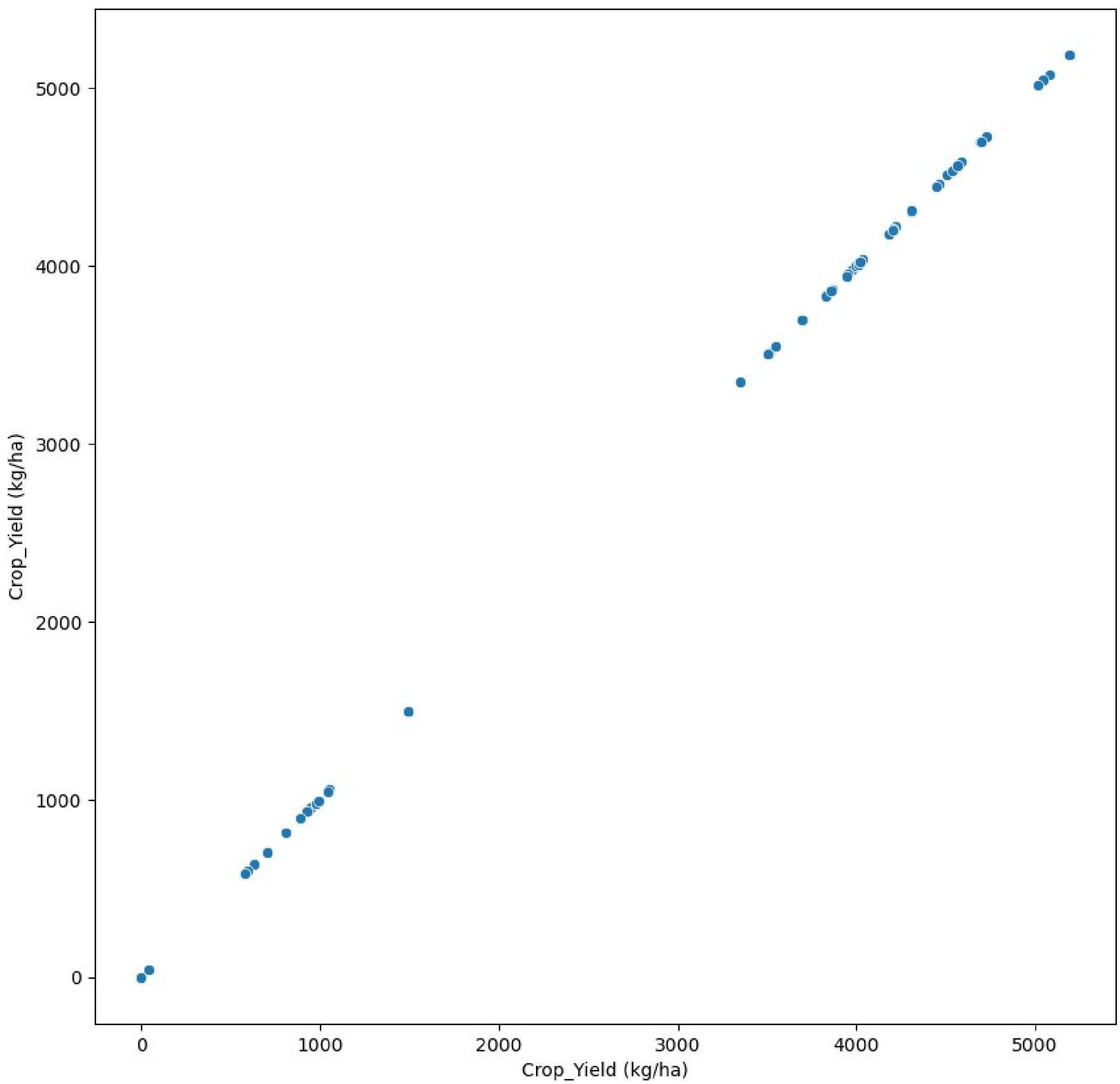

```
for i in z.columns:
    if(z[i].dtype != "object"):
        plt.figure(figsize = (10, 10))
        sns.scatterplot(x = z[i], y = z["Crop_Yield (kg/ha)"], data =
z)
```



Crop_Yield (kg/ha)
0
1000
2000
3000
4000
5000
2000
2005
2010
2015
2020
Year

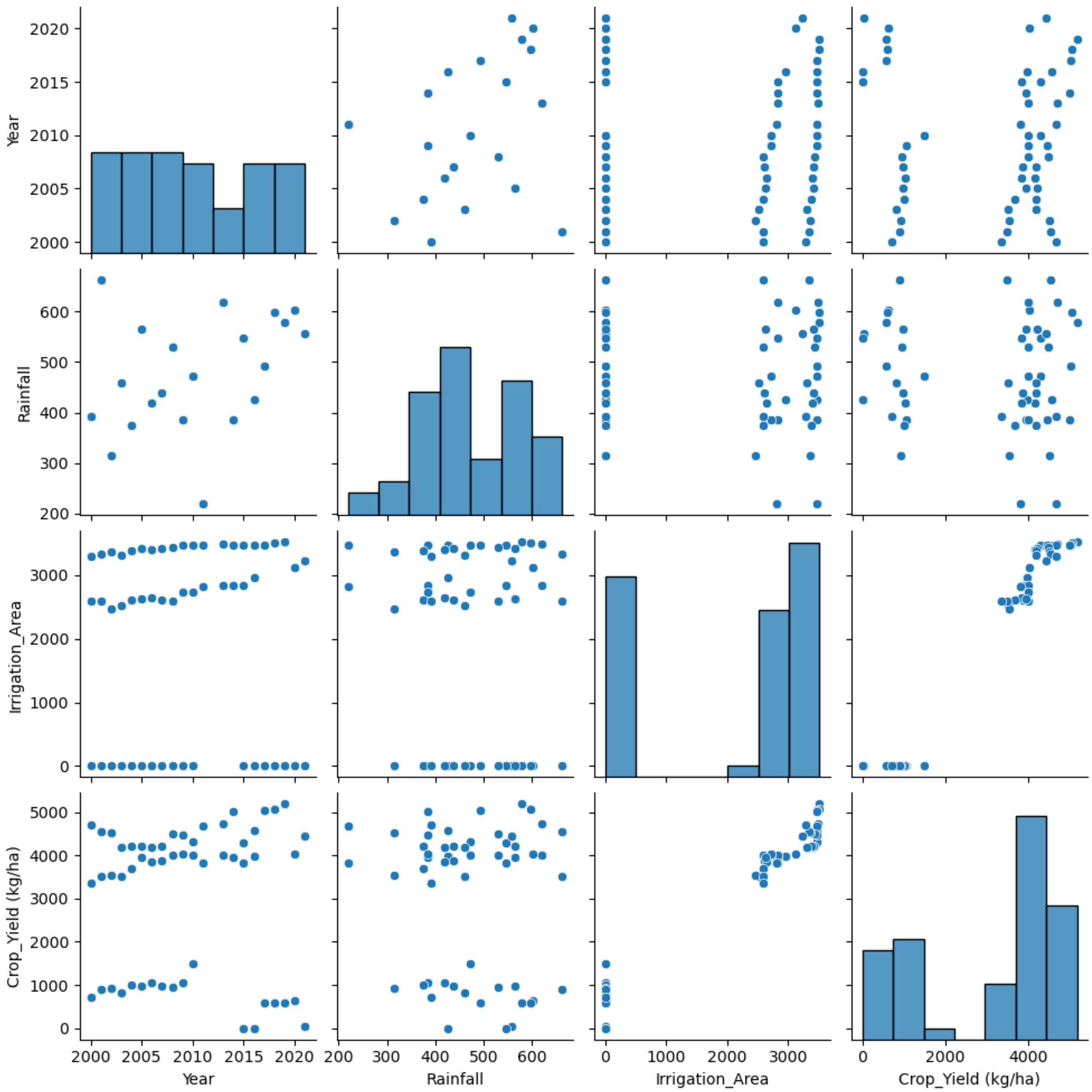

5000
4000
3000
2000
1000
0
Crop_Yield (kg/ha)
200
300
400
500
600
Rainfall


Crop_Yield (kg/ha)
Irrigation_Area
0
1000
2000
3000
4000
5000
0
500
1000
1500
2000
2500
3000
3500




```
sns.pairplot(z)
```

```
<seaborn.axisgrid.PairGrid at 0x1cb95192a80>
```




```
z.columns
```

```
Index(['id', 'Year', 'State', 'Crop_Type', 'Rainfall', 'Soil_Type',
       'Irrigation_Area', 'Crop_Yield (kg/ha)'],
      dtype='object')
```

```
pd.DataFrame(z.groupby(["State", "Crop_Type", "Soil_Type"])
["Crop_Yield (kg/ha)"].sum())
```

```
                            Crop_Yield (kg/ha)
State  Crop_Type Soil_Type
Punjab Bajra     Loamy                   13263
```

Rice	alluvial	69389
Wheat	Loamy	86716

```
b = z.copy()
for i in b.columns:
    if(b[i].dtype == "object"):
        b.drop([i], axis = 1, inplace = True)
b
```

	Year	Rainfall	Irrigation_Area	Crop_Yield (kg/ha)
0	2019	578.6	3515.2	5188
1	2018	598.3	3499.3	5077
2	2017	493.0	3467.7	5046
3	2016	426.7	3474.6	4583
4	2015	546.9	3474.7	4304
5	2014	384.9	3474.7	5017
6	2013	619.7	3488.1	4724
7	2011	218.9	3466.9	4693
8	2010	472.1	3474.8	4307
9	2009	384.9	3474.8	4462
10	2008	529.2	3437.9	4507
11	2007	438.0	3406.9	4210
12	2006	418.3	3404.8	4179
13	2005	565.9	3410.5	4221
14	2004	375.2	3381.7	4207
15	2003	459.5	3311.6	4200
16	2002	314.5	3353.5	4532
17	2001	662.8	3333.6	4563
18	2000	391.9	3284.3	4696
19	2021	556.9	3229.5	4443
20	2020	602.6	3118.8	4034
21	2016	426.7	2961.4	3974
22	2015	546.9	2838.3	3838
23	2014	384.9	2838.3	3952
24	2013	619.7	2837.6	3998
25	2011	218.9	2814.2	3828
26	2010	472.1	2721.8	4010
27	2009	384.9	2721.8	4022
28	2008	529.2	2592.2	4019
29	2007	438.0	2602.4	3868
30	2006	418.3	2639.9	3858
31	2005	565.9	2632.3	3943
32	2004	375.2	2599.6	3694
33	2003	459.5	2515.7	3510
34	2002	314.5	2471.0	3545
35	2001	662.8	2590.3	3506
36	2000	391.9	2584.7	3347
37	2021	556.9	3.9	40
38	2020	602.6	2.0	635
39	2019	578.6	1.9	583

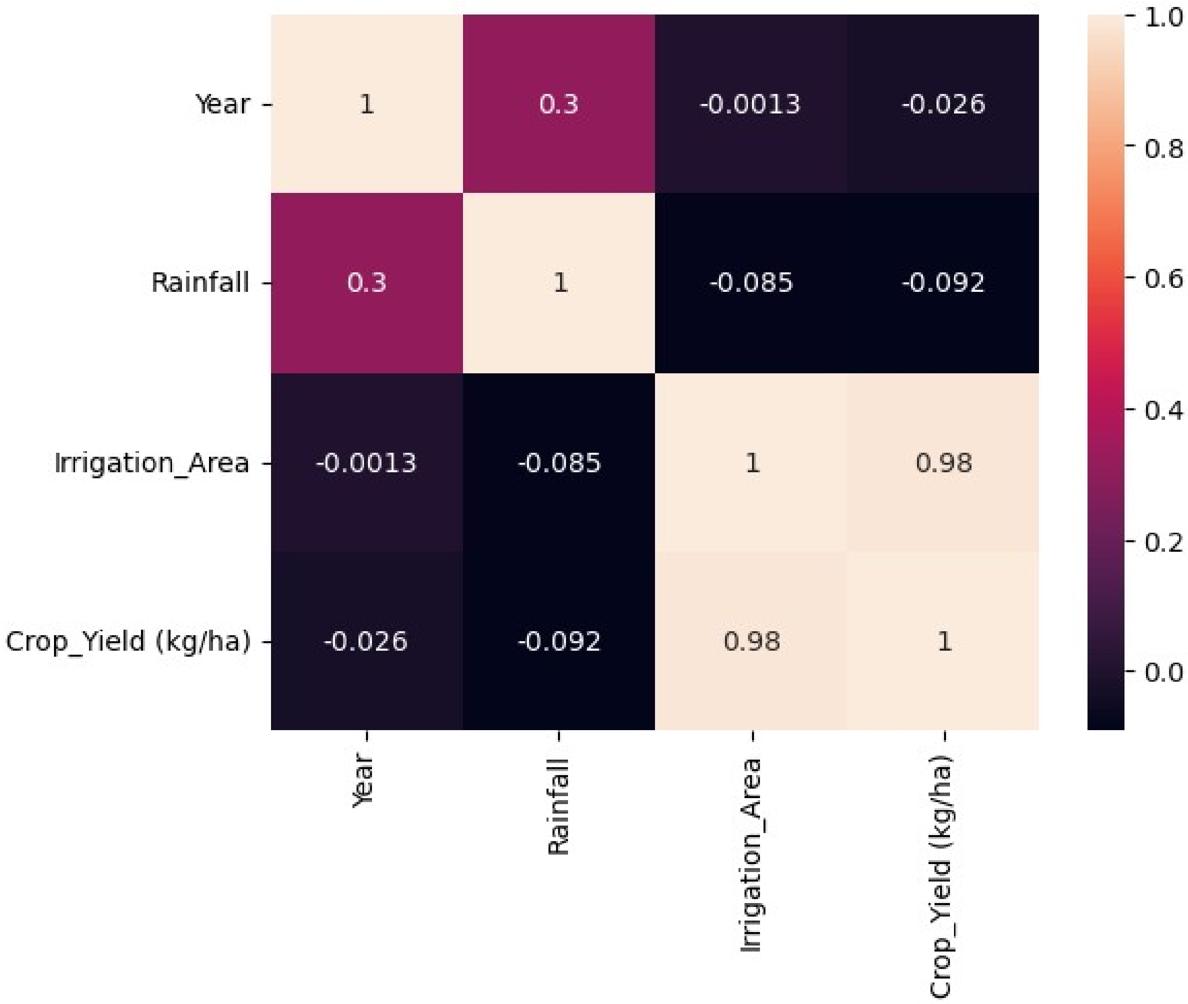
```
40  2018     598.3              2.8                 597
41  2017     493.0              3.1                 580
42  2016     426.7              1.9                   0
43  2015     546.9              1.2                   0
44  2010     472.1              4.9                1495
45  2009     384.9              4.9                1055
46  2008     529.2              3.5                 950
47  2007     438.0              4.2                 977
48  2006     418.3              5.2                1045
49  2005     565.9              5.6                 978
50  2004     375.2              7.2                 993
51  2003     459.5              6.1                 810
52  2002     314.5              7.6                 929
53  2001     662.8              5.4                 893
54  2000     391.9              4.6                 703
```

```
b.corr()
```

```
                        Year  Rainfall  Irrigation_Area  Crop_Yield
(kg/ha)
Year                1.000000  0.304973        -0.001326            -
0.026250
Rainfall            0.304973  1.000000        -0.085025            -
0.092148
Irrigation_Area    -0.001326 -0.085025         1.000000
0.984287
Crop_Yield (kg/ha) -0.026250 -0.092148         0.984287
1.000000
```

```
sns.heatmap(b.corr(), annot = True)
```

```
<Axes: >
```



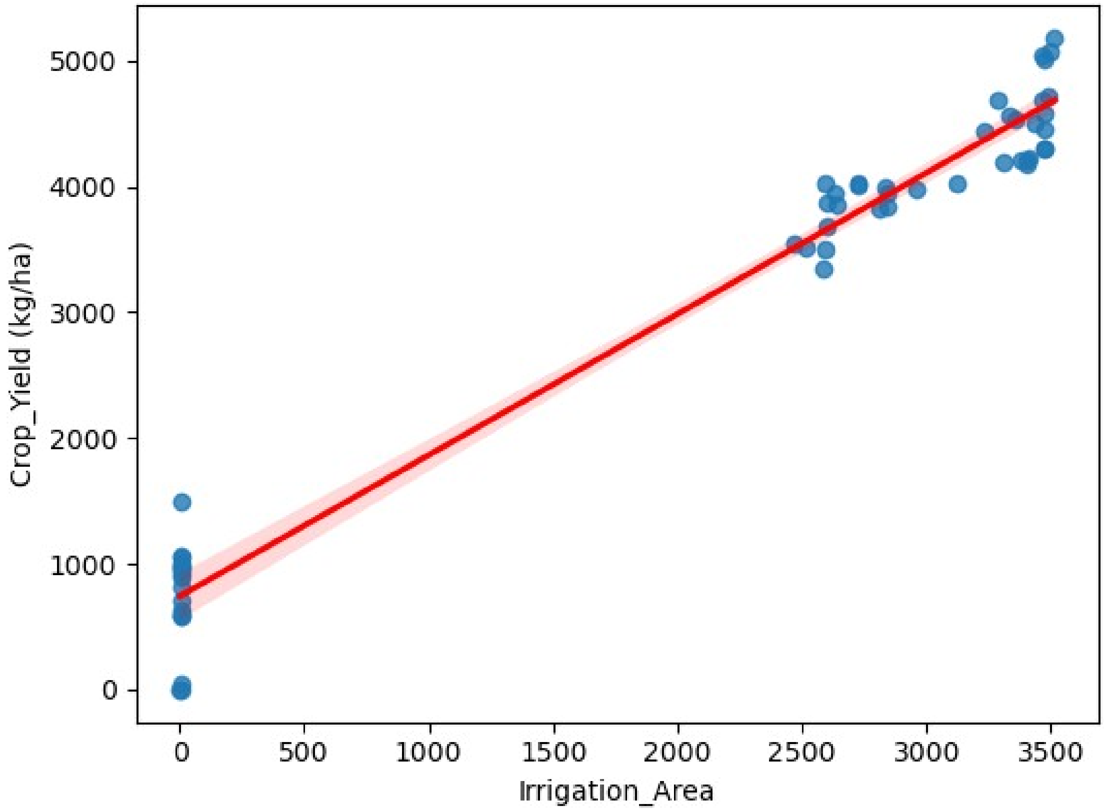

```
b.corr()["Crop_Yield (kg/ha)"].sort_values(ascending = False)
```

```
Crop_Yield (kg/ha)     1.000000
Irrigation_Area        0.984287
Year                  -0.026250
Rainfall              -0.092148
Name: Crop_Yield (kg/ha), dtype: float64
```

Regression analysis

```
sns.regplot(x = z["Irrigation_Area"], y = z["Crop_Yield (kg/ha)"],
data = z, line_kws = {"color" : "red"})
```

```
<Axes: xlabel='Irrigation_Area', ylabel='Crop_Yield (kg/ha)'>
```



Model selection

```
X = b["Irrigation_Area"]
Y = b["Crop_Yield (kg/ha)"]

x_train, x_test, y_train, y_test = train_test_split(X, Y, train_size =
0.7, test_size = 0.3, random_state = 100)
```

Reshaping x_train

```
x_train = np.array(x_train).reshape(-1, 1)
```

Reshaping y_train

```
y_train = np.array(y_train).reshape(-1, 1)
```

Training model using train dataset

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

LinearRegression()
```

Evaluation of Training dataset

```
y_predict_train = n.predict(x_train)
r2_train = r2_score(y_true = y_train, y_pred = y_predict_train)

round(r2_train, 2)*100

97.0

mse_train = mse(y_true = y_train, y_pred = y_predict_train)

rmse_train = np.sqrt(mse_train)

rmse_train

294.06160727273607
```

Reshaping x_test

```
x_test = np.array(x_test).reshape(-1, 1)
```

Reshaping y_test

```
y_test = np.array(y_test).reshape(-1, 1)
```

Evaluation of Testing dataset

```
y_predict_test = n.predict(x_test)
r2_test = r2_score(y_true = y_test, y_pred = y_predict_test)

round(r2_test, 2)*100

97.0

mse_test = mse(y_true = y_test, y_pred = y_predict_test)

rmse_test = np.sqrt(mse_test)
rmse_test

320.86173720697604
```

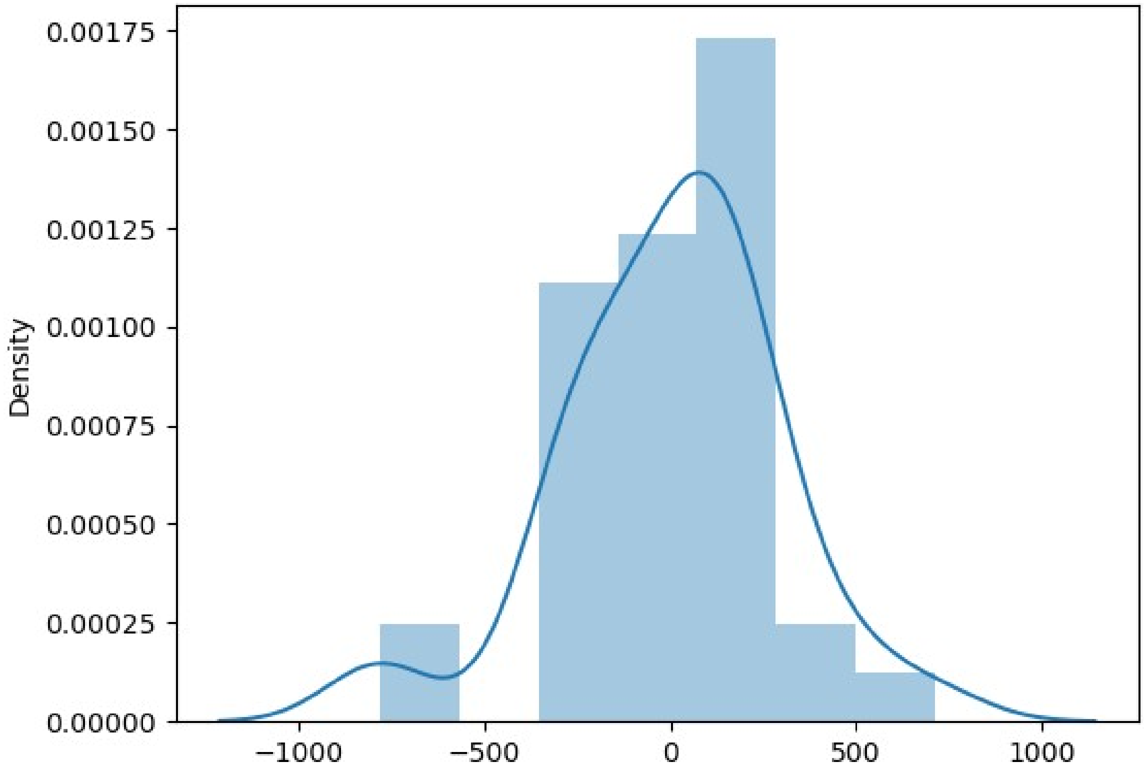
Residual analysis for training dataset

```
res_train = y_train - y_predict_train

res_train

array([[-202.3891355 ],
       [-781.07631834],
       [-106.82293029],
       [  24.3288216 ],
       [ 136.99624114],
       [ 121.16413507],
```

```
       [ 141.68780015],
       [ -29.77273187],
       [ 167.17325878],
       [ 130.88312969],
       [ -44.81163671],
       [-289.13345846],
       [ 469.71807177],
       [-276.36917647],
       [-324.89757696],
       [ 265.31418937],
       [ 710.64163876],
       [ -46.13847648],
       [ 241.93949669],
       [-118.4785779 ],
       [  84.22535659],
       [-259.69645021],
       [ 192.87582875],
       [ 108.09463161],
       [-780.31050833],
       [-271.62584665],
       [ 270.64163876],
       [ 473.28892363],
       [  62.6929756 ],
       [ 190.91396047],
       [ -33.10945026],
       [ 260.31343447],
       [-201.93544427],
       [  67.86152352],
       [-198.07631834],
       [   2.74022496],
       [ 114.62733353],
       [-273.4785779 ]])
```

```
sns.distplot(res_train, kde = True)
```

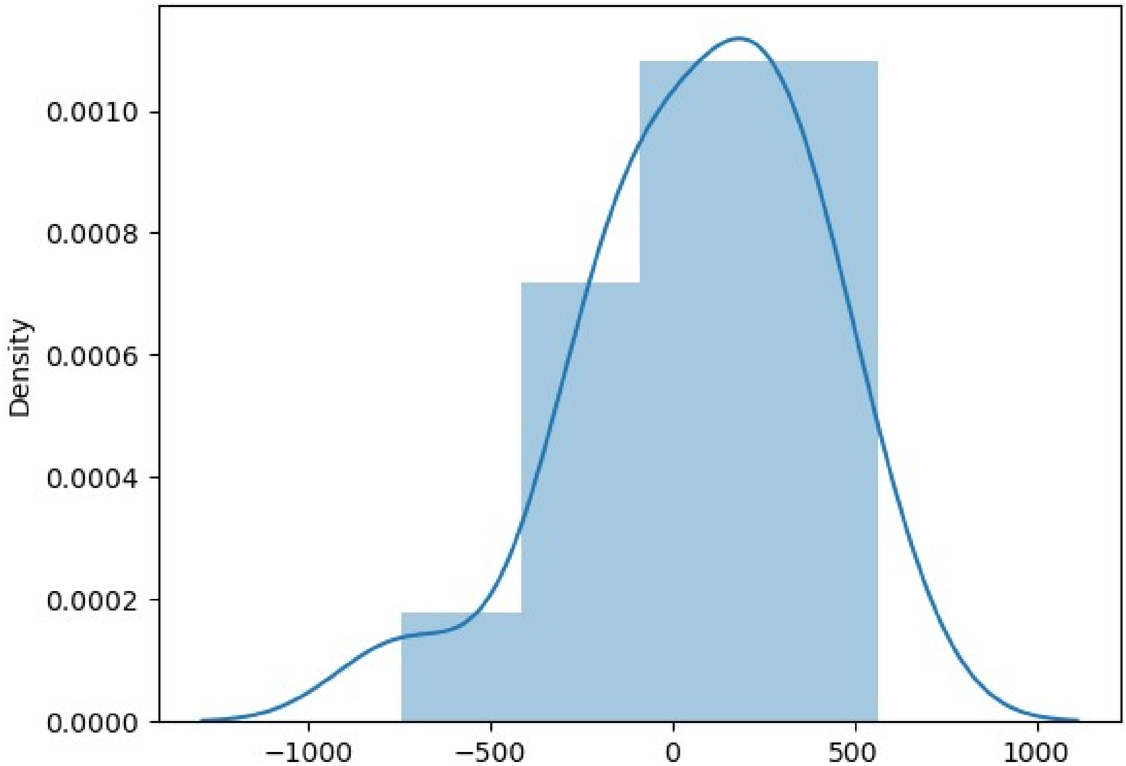
```
<Axes: ylabel='Density'>
```



Residual analysis for testing dataset

```
res_test = y_test - y_predict_test

res_test

array([[ 404.09844254],
       [ 284.22846915],
       [  71.00273673],
       [-296.19500698],
       [-185.06093121],
       [ 193.40744877],
       [ 563.32324442],
       [ 128.97103192],
       [ 253.31418937],
       [ -81.03015695],
       [ 436.63082353],
       [-157.0094874 ],
       [ 206.12540587],
       [ 323.93114609],
       [-146.18571977],
       [ -21.20946357],
       [-743.26434694]])

sns.distplot(res_test, kde = True)
```

```
<Axes: ylabel='Density'>
```



Recommendations

Crop yield prediction models play a pivotal role in modern agriculture, bridging the gap between data analysis and actionable insights. By leveraging comprehensive datasets and advanced machine learning techniques, stakeholders can drive sustainable growth, improve resource utilization, and secure food supplies for the future. This analysis underscores the transformative potential of data-driven approaches in addressing global agricultural challenges.