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

In [2]:
```
df1 = pd.read_csv('AllOdiPlayers.csv')
df2 = pd.read_csv('AllOdicaptains.csv')
df3 = pd.read_csv('AllTestcaptains.csv')
df4 = pd.read_csv('AllTestplayers.csv')
```

In [3]:
```
df1.head(10)
```

Out[3]:

	No	Name	First	Last	Mat	Inn	NO	Runs	HS	Avg	Balls	Mdn	Runs.1	Wkt	BB
0	1	Syed Abid Ali	1974	1975	5	3	0	93	70	31.00	336	10	187	7	2 Fe
1	2	Bishan Singh Bedi	1974	1979	10	7	2	31	13	6.20	590	17	340	7	Fe
2	3	Farokh Engineer	1974	1975	5	4	1	114	54*	38.00	0	0	0	0	
3	4	Sunil Gavaskar	1974	1987	108	102	14	3092	103*	35.13	20	0	25	1	1 Ja
4	5	Madan Lal	1974	1987	67	35	14	401	53*	19.09	3164	44	2137	73	2 A
5	6	Sudhir Naik	1974	1974	2	2	0	38	20	19.00	0	0	0	0	
6	7	Brijesh Patel	1974	1979	10	9	1	243	82	30.37	0	0	0	0	
7	8	Eknath Solkar	1974	1976	7	6	0	27	13	4.50	252	4	169	4	Fe
8	9	Srinivasaraghavan Venkataraghavan	1974	1983	15	9	4	54	26*	10.80	868	7	542	5	Fe
9	10	Gundappa Viswanath	1974	1982	25	23	1	439	75	19.95	0	0	0	0	

In [4]:
```
df1.tail(5)
```

Out[4]:

	No	Name	First	Last	Mat	Inn	NO	Runs	HS	Avg	Balls	Mdn	Runs.1	Wkt	BBM	Av
236	237	Rahul Chahar	2021	2021	1	1	0	13	13	13.0	60	0	54	3	Mar-54	
237	238	Krishnappa Gowtham	2021	2021	1	1	0	2	2	2.0	48	0	49	1	Jan-49	
238	239	Nitish Rana	2021	2021	1	1	0	7	7	7.0	18	0	10	0	0	

	No	Name	First	Last	Mat	Inn	NO	Runs	HS	Avg	Balls	Mdn	Runs.1	Wkt	BBM	Av
239	240	Chetan Sakariya	2021	2021	1	1	0	0	0*	0.0	48	0	34	2	Feb-34	
240	241	Sanju Samson	2021	2021	1	1	0	46	46	46.0	0	0	0	0	0	

In [5]:
```
df1.columns
```

Out[5]:
```
Index(['No', 'Name', 'First', 'Last', 'Mat', 'Inn', 'NO', 'Runs', 'HS', 'Avg',
       'Balls', 'Mdn', 'Runs.1', 'Wkt', 'BBM', 'Avg.1', 'Ca', 'St'],
      dtype='object')
```

In [6]:
```
df1 =df1.drop('BBM', axis =1)
```

In [7]:
```
df1 = df1.drop('Mdn', axis = 1)
df1 = df1.drop('Runs.1', axis = 1)
df1 = df1.drop('Wkt', axis = 1)
df1 = df1.drop('Avg.1', axis = 1)
df1 = df1.drop('Ca', axis = 1)
df1 = df1.drop('St', axis = 1)
df1 = df1.drop('Balls', axis = 1)
df1 = df1.drop('NO', axis = 1)
df1 = df1.drop('No', axis = 1)
df1 = df1.drop('First', axis = 1)
df1 = df1.drop('Last', axis = 1)
```

In [8]:
```
df1
```

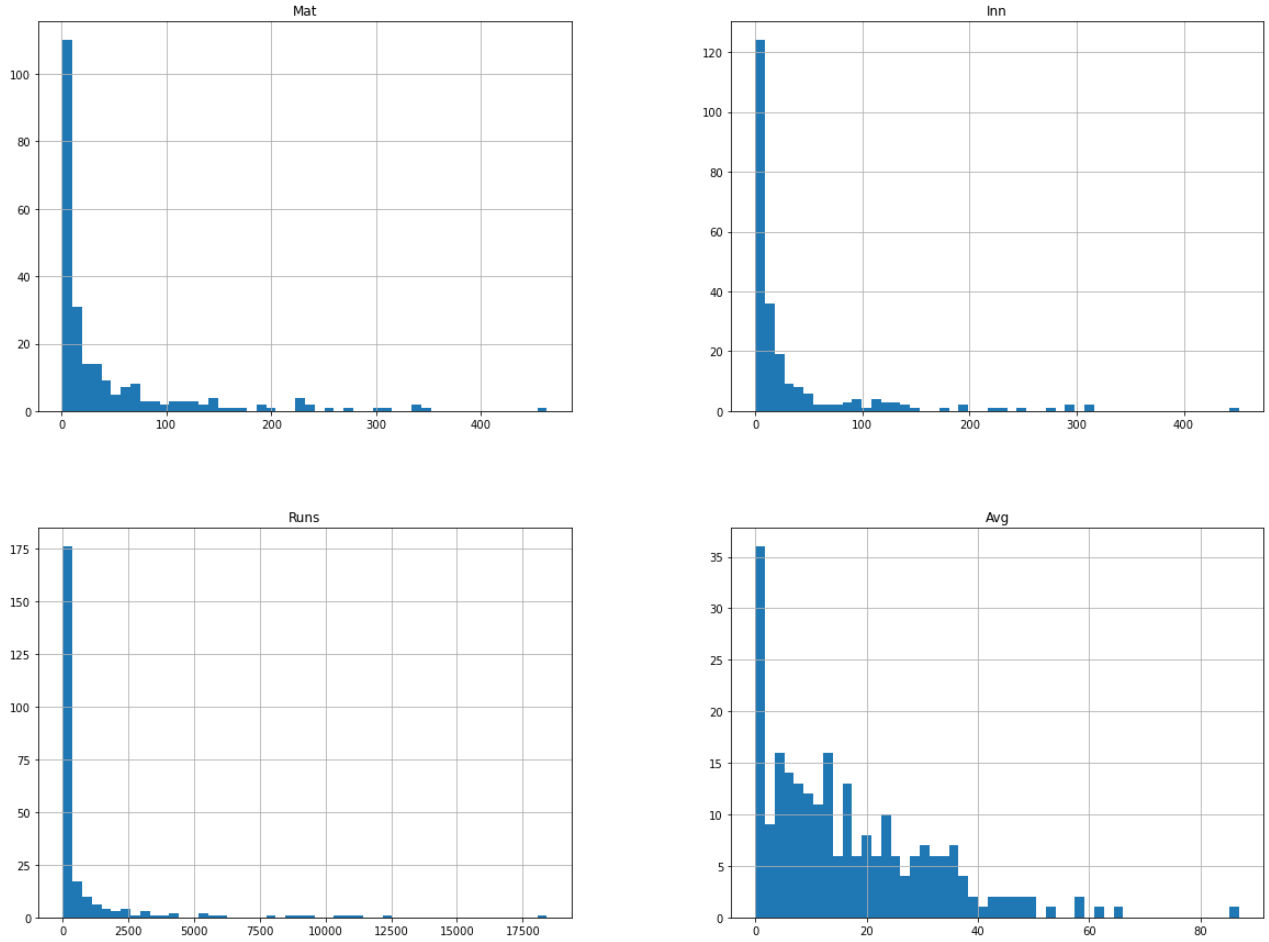
Out[8]:

	Name	Mat	Inn	Runs	HS	Avg
0	Syed Abid Ali	5	3	93	70	31.00
1	Bishan Singh Bedi	10	7	31	13	6.20
2	Farokh Engineer	5	4	114	54*	38.00
3	Sunil Gavaskar	108	102	3092	103*	35.13
4	Madan Lal	67	35	401	53*	19.09
...	...	...	...	...	...	...
236	Rahul Chahar	1	1	13	13	13.00
237	Krishnappa Gowtham	1	1	2	2	2.00
238	Nitish Rana	1	1	7	7	7.00
239	Chetan Sakariya	1	1	0	0*	0.00
240	Sanju Samson	1	1	46	46	46.00

241 rows × 6 columns

In [9]:
```
df1.hist(bins=50,figsize =(20,15))
plt.show
```

Out[9]:
```
<function matplotlib.pyplot.show(close=None, block=None)>
```



In [10]:
```
df1.isna().sum()
```

Out[10]:
```
Name    0
Mat     0
Inn     0
Runs    0
HS      0
Avg     0
dtype: int64
```

In [11]:
```
df1 = df1.drop('Name', axis = 1)
```

In [12]:
```
df1 = df1.drop('HS', axis = 1)
```

In [13]:
```
from sklearn import datasets, linear_model, metrics
X = df1
y = df1.Runs
```

In [14]:
```
from sklearn.model_selection import train_test_split
```

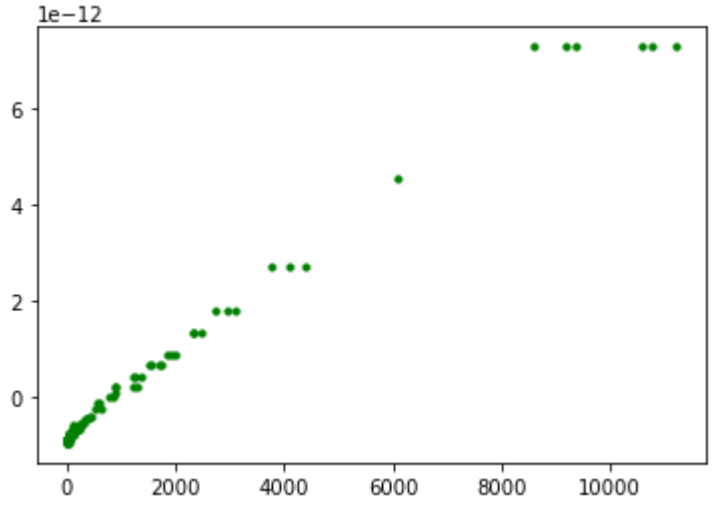
```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.4,random_state=1)
```

In [15]:
```
from sklearn.linear_model import LinearRegression
lin_regx = LinearRegression()
lin_regx.fit(X_train, y_train)
```

Out[15]:
```
LinearRegression()
```

In [16]:
```
plt.scatter(lin_regx.predict(X_train), lin_regx.predict(X_train) - y_train,
            color = "green", s = 10, label = 'Train data')
```

Out[16]:
```
<matplotlib.collections.PathCollection at 0x228e77c1d90>
```



In [ ]:
```

```