In [1]:

```
import pandas as pd
import matplotlib.pyplot as plot
%matplotlib inline
from copy import deepcopy
import numpy as np
import seaborn as sns
sns.set()
from matplotlib import pyplot as plt
```

In [2]:

```
df=pd.read_csv("Desktop/dataset_icc.csv")
```

In [3]:

```
df.head(10)
```

Out[3]:

	Player	Span	Mat	Inn	NO	Runs	HS	Avg	100	50	0	
0	SR Tendulkar+//0AoA-(INDIA)	1989+AC0-2013	200	329	33	15921	248+ACo-	53.78	51	68	14	http://stats
1	RT Ponting+//0AoA-(AUS)	1995+AC0-2012	168	287	29	13378	257	51.85	41	62	17	http://stats
2	JH Kallis+//0AoA-(ICC/SA)	1995+AC0-2013	166	280	40	13289	224	55.37	45	58	16	http://stats
3	R Dravid+//0AoA-(ICC/INDIA)	1996+AC0-2012	164	286	32	13288	270	52.31	36	63	8	http://stats
4	AN Cook+//0AoA-(ENG)	2006+AC0-2018	161	291	16	12472	294	45.35	33	57	9	http://stats
5	KC Sangakkara+//0AoA-(SL)	2000+AC0-2015	134	233	17	12400	319	57.40	38	52	11	http://stats
6	BC Lara+//0AoA-(ICC/WI)	1990+AC0-2006	131	232	6	11953	400+ACo-	52.88	34	48	17	http://stats
7	S Chanderpaul+//0AoA-(WI)	1994+AC0-2015	164	280	49	11867	203+ACo-	51.37	30	66	15	http://stats
8	DPMD Jayawardene+//0AoA-(SL)	1997+AC0-2014	149	252	15	11814	374	49.84	34	50	15	http://stats
9	AR Border+//0AoA-(AUS)	1978+AC0-1994	156	265	44	11174	205	50.56	27	63	11	http://stats

In [4]:

```
print("No of players in the dataset: " +str(len(df.index)))
```

```
No of players in the dataset: 1476
```

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

```
               Mat          Inn           NO          Runs          Avg  \
count  1476.000000  1476.000000  1476.000000   1476.000000  1476.000000   
mean     31.393631    51.126016     5.932249   1454.179539    28.323076   
std      29.330591    47.689011     7.713920   1974.937261    12.979955   
min       2.000000     3.000000     0.000000    188.000000     4.760000   
25%      11.000000    18.000000     1.000000    324.500000    18.880000   
50%      21.000000    34.000000     3.000000    682.500000    26.905000   
75%      41.250000    68.000000     8.000000   1661.250000    35.697500   
max     200.000000   329.000000    89.000000  15921.000000   160.500000   

               100           50            0  
count  1476.000000  1476.000000  1476.000000  
mean      2.834011     6.808266     5.011518  
std       5.713993     9.784067     5.006356  
min       0.000000     0.000000     0.000000  
25%       0.000000     1.000000     1.000000  
50%       1.000000     3.000000     4.000000  
75%       3.000000     8.000000     7.000000  
max      51.000000    68.000000    43.000000  
```

In [6]:
```
print(df.shape)
```

```
(1476, 12)
```

In [7]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1476 entries, 0 to 1475
Data columns (total 12 columns):
Player            1476 non-null object
Span              1476 non-null object
Mat               1476 non-null int64
Inn               1476 non-null int64
NO                1476 non-null int64
Runs              1476 non-null int64
HS                1476 non-null object
Avg               1476 non-null float64
100               1476 non-null int64
50                1476 non-null int64
0                 1476 non-null int64
Player Profile    1476 non-null object
dtypes: float64(1), int64(7), object(4)
memory usage: 138.5+ KB
```

```
In [8]: df.isnull()
```

```
Out[8]:
        Player  Span   Mat   Inn    NO  Runs    HS   Avg   100    50     0  Player Profile
   0     False False False False False False False False False False False           False
   1     False False False False False False False False False False False           False
   2     False False False False False False False False False False False           False
   3     False False False False False False False False False False False           False
   4     False False False False False False False False False False False           False
 ...       ...   ...   ...   ...   ...   ...   ...   ...   ...   ...   ...             ...
1471     False False False False False False False False False False False           False
1472     False False False False False False False False False False False           False
1473     False False False False False False False False False False False           False
1474     False False False False False False False False False False False           False
1475     False False False False False False False False False False False           False

1476 rows × 12 columns
```

```
In [12]: df.drop("Player Profile",axis=1,inplace=True)
```

```
In [15]: df.drop("Span",axis=1,inplace=True)
```

In [16]:

```
df.head(20)
```

Out[16]:

	Player	Mat	Inn	NO	Runs	HS	Avg	100	50	0
0	SR Tendulkar+/0AoA-(INDIA)	200	329	33	15921	248+ACo-	53.78	51	68	14
1	RT Ponting+//0AoA-(AUS)	168	287	29	13378	257	51.85	41	62	17
2	JH Kallis+//0AoA-(ICC/SA)	166	280	40	13289	224	55.37	45	58	16
3	R Dravid+//0AoA-(ICC/INDIA)	164	286	32	13288	270	52.31	36	63	8
4	AN Cook+//0AoA-(ENG)	161	291	16	12472	294	45.35	33	57	9
5	KC Sangakkara+//0AoA-(SL)	134	233	17	12400	319	57.40	38	52	11
6	BC Lara+//0AoA-(ICC/WI)	131	232	6	11953	400+ACo-	52.88	34	48	17
7	S Chanderpaul+//0AoA-(WI)	164	280	49	11867	203+ACo-	51.37	30	66	15
8	DPMD Jayawardene+//0AoA-(SL)	149	252	15	11814	374	49.84	34	50	15
9	AR Border+//0AoA-(AUS)	156	265	44	11174	205	50.56	27	63	11
10	SR Waugh+//0AoA-(AUS)	168	260	46	10927	200	51.06	32	50	22
11	SM Gavaskar+//0AoA-(INDIA)	125	214	16	10122	236+ACo-	51.12	34	45	12
12	Younis Khan+//0AoA-(PAK)	118	213	19	10099	313	52.05	34	33	19
13	HM Amla+//0AoA-(SA)	124	215	16	9282	311+ACo-	46.64	28	41	13
14	GC Smith+//0AoA-(ICC/SA)	117	205	13	9265	277	48.25	27	38	11
15	GA Gooch+//0AoA-(ENG)	118	215	6	8900	333	42.58	20	46	13
16	Javed Miandad+//0AoA-(PAK)	124	189	21	8832	280+ACo-	52.57	23	43	6
17	Inzamam+AC0-ul+AC0-Haq+//0AoA-(ICC/PAK)	120	200	22	8830	329	49.60	25	46	15
18	VVS Laxman+//0AoA-(INDIA)	134	225	34	8781	281	45.97	17	56	14
19	AB de Villiers+//0AoA-(SA)	114	191	18	8765	278+ACo-	50.66	22	46	8

In [18]:

```
import numpy as np
import seaborn as sns
sns.set()   #for plot styling
from matplotlib import pyplot as plt
```

In [21]:
```
df['Avg'].head(10)
```

Out[21]:
```
0    53.78
1    51.85
2    55.37
3    52.31
4    45.35
5    57.40
6    52.88
7    51.37
8    49.84
9    50.56
Name: Avg, dtype: float64
```

In [22]:
```
df['Mat'].head(10)
```

Out[22]:
```
0    200
1    168
2    166
3    164
4    161
5    134
6    131
7    164
8    149
9    156
Name: Mat, dtype: int64
```

In [23]:
```
c_data=df.iloc[:,7:9]
c_data
```

Out[23]:

	100	50
0	51	68
1	41	62
2	45	58
3	36	63
4	33	57
...	...	...
1471	0	0
1472	0	0
1473	0	1
1474	1	0
1475	0	1

1476 rows × 2 columns

In [24]:
```
from sklearn.cluster import KMeans
```

```
In [25]: kmeans=KMeans(n_clusters=5)
```

```
In [27]: print(c_data)
```

```
      100  50
0      51  68
1      41  62
2      45  58
3      36  63
4      33  57
...   ...  ..
1471    0   0
1472    0   0
1473    0   1
1474    1   0
1475    0   1

[1476 rows x 2 columns]
```

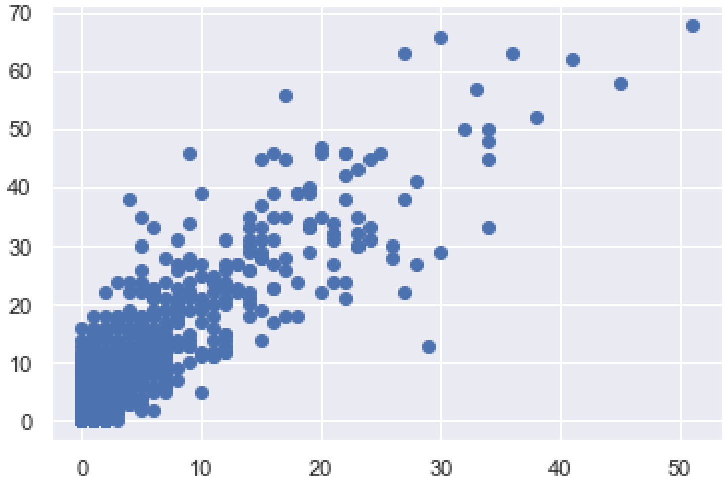
```
In [28]: x=np.array(c_data)
```

```
In [29]: print(x)
```

```
[[51 68]
 [41 62]
 [45 58]
 ...
 [ 0  1]
 [ 1  0]
 [ 0  1]]
```

```
In [30]: plt.scatter(x[:,0],x[:,1],label='True Position')
```

```
Out[30]: <matplotlib.collections.PathCollection at 0xbfba5c8>
```

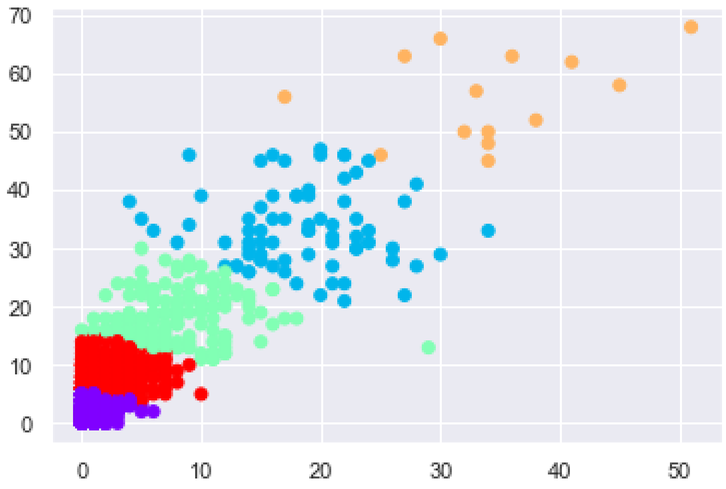


```
In [31]: kmeans.fit(x)

Out[31]: KMeans(algorithm='auto', copy_x=True, init='k-means++', max_iter=300,
                n_clusters=5, n_init=10, n_jobs=None, precompute_distances='auto',
                random_state=None, tol=0.0001, verbose=0)
```

```
In [32]: kmeans.cluster_centers_

Out[32]: array([[ 0.41675504,  1.73913043],
                [18.24324324, 33.41891892],
                [ 7.88666667, 18.20666667],
                [34.07142857, 56.        ],
                [ 2.6440678 ,  8.20677966]])
```

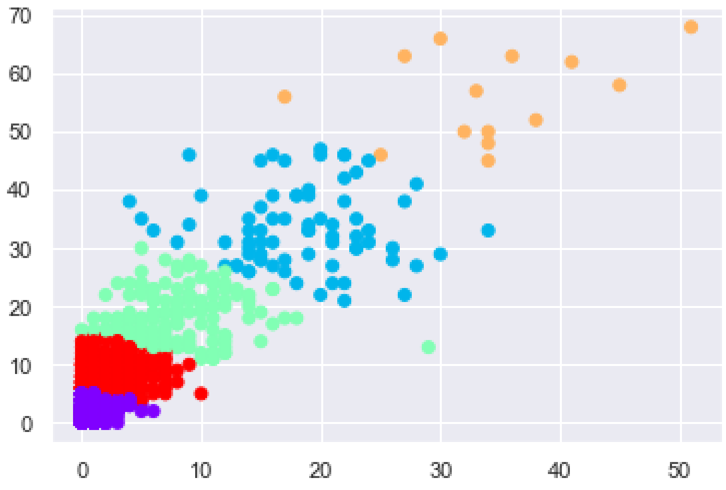
```
In [33]: kmeans.labels_

Out[33]: array([3, 3, 3, ..., 0, 0, 0])
```

```
In [35]: plt.scatter(x[ : , 0], x[ : , 1],c=kmeans.labels_,cmap='rainbow')

Out[35]: <matplotlib.collections.PathCollection at 0xc2cd048>
```



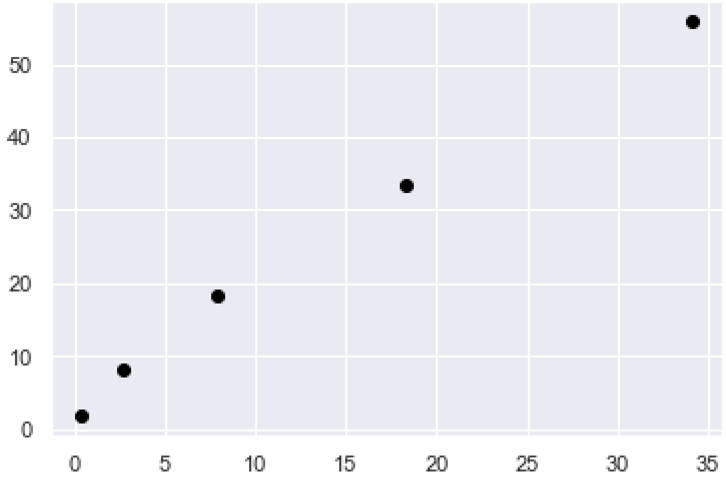
In [36]:
```
plt.scatter(x[ : , 0], x[ : , 1],c=kmeans.labels_,cmap='rainbow')
```

Out[36]:
```
<matplotlib.collections.PathCollection at 0xc339ac8>
```



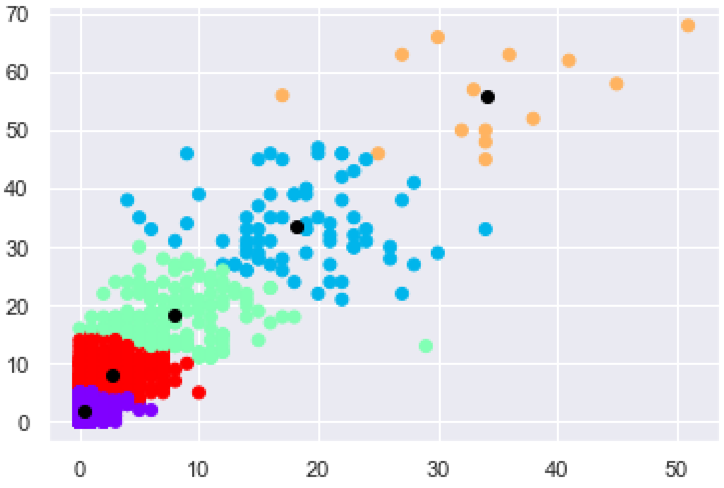
In [37]:
```
plt.scatter(kmeans.cluster_centers_[:,0],kmeans.cluster_centers_[:,1],color='bla
```

Out[37]:
```
<matplotlib.collections.PathCollection at 0xc3a2ac8>
```

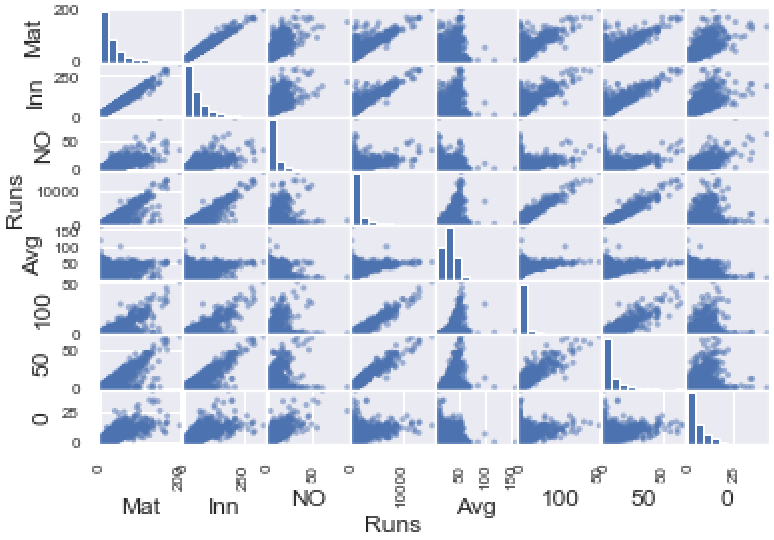


In [38]:
```
plt.scatter(x[ : , 0], x[ : , 1],c=kmeans.labels_,cmap='rainbow')
plt.scatter(kmeans.cluster_centers_[:,0],kmeans.cluster_centers_[:,1],color='bla
```

Out[38]:
```
<matplotlib.collections.PathCollection at 0xc410248>
```

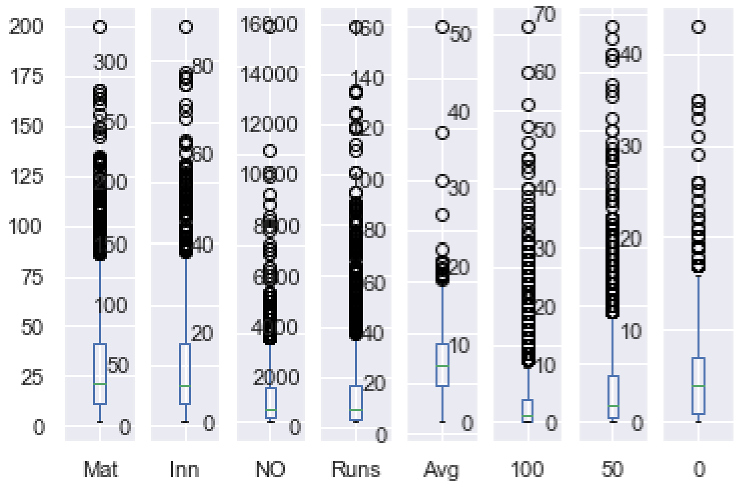


In [40]:
```
from pandas.plotting import scatter_matrix
scatter_matrix(df)
plt.show()
```

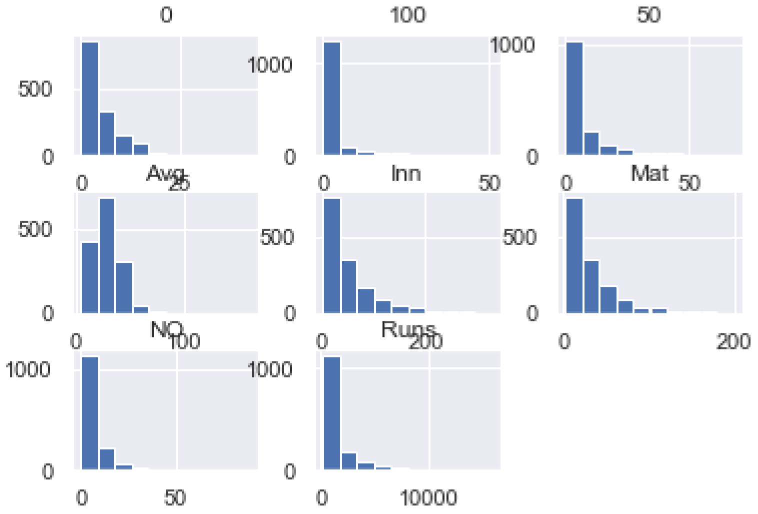


In [45]:
```
df.plot(kind='box', subplots=True, sharex=False, sharey=False )
```

Out[45]:
```
Mat        AxesSubplot(0.125,0.125;0.0824468x0.755)
Inn     AxesSubplot(0.223936,0.125;0.0824468x0.755)
NO      AxesSubplot(0.322872,0.125;0.0824468x0.755)
Runs    AxesSubplot(0.421809,0.125;0.0824468x0.755)
Avg     AxesSubplot(0.520745,0.125;0.0824468x0.755)
100     AxesSubplot(0.619681,0.125;0.0824468x0.755)
50      AxesSubplot(0.718617,0.125;0.0824468x0.755)
0       AxesSubplot(0.817553,0.125;0.0824468x0.755)
dtype: object
```



In [46]:
```
df.hist()
plt.show()
```



In [47]:
```
from sklearn import metrics
```

In [50]:
```
from sklearn.metrics import f1_score,recall_score,precision_score
```

In [52]:

```
print(metrics.confusion_matrix(x[:,0],x[:,1]))
```

```
[[198 182 143 ...   0   0   0]
 [ 38  39  36 ...   0   0   0]
 [  7   7  11 ...   0   0   0]
 ...
 [  0   0   0 ...   0   0   0]
 [  0   0   0 ...   0   0   0]
 [  0   0   0 ...   0   0   0]]
```

In [ ]: