In [1]:
```
import pandas as pd

# URL of the CSV file
url = 'https://github.com/ghoshaustin/My_Public_Dataset_Notebook/blob/main/cricket.csv?raw=true'

# Read the CSV file directly from the URL
df = pd.read_csv(url)

# Display the first few rows of the dataframe
df.head(5)
```

Out[1]:

	Player	Span	Mat	Inns	NO	Runs	HS	Ave	BF	SR	100	50	0	4s	6s
0	DG Bradman (AUS)	1928-1948	52	80	10	6996	334	99.94	9800+	58.60	29	13	7	626+	6
1	HC Brook (ENG)	2022-2023	12	20	1	1181	186	62.15	1287	91.76	4	7	1	141	23
2	AC Voges (AUS)	2015-2016	20	31	7	1485	269*	61.87	2667	55.68	5	4	2	186	5
3	RG Pollock (SA)	1963-1970	23	41	4	2256	274	60.97	1707+	54.48	7	11	1	246+	11
4	GA Headley (WI)	1930-1954	22	40	4	2190	270*	60.83	416+	56.00	10	5	2	104+	1

In [2]:
```
# rename col
df = df.rename(columns = {'Mat': 'Matches','NO': 'Not_Out', 'HS': 'Highest_in_score', 'SR': 'Batting_strike_rate' })
```

In [3]:
```
df.head(2)
```

Out[3]:

	Player	Span	Matches	Inns	Not_Out	Runs	Highest_in_score	Ave	BF	Batting_strike_rate	100	50	0	4s	6s
0	DG Bradman (AUS)	1928-1948	52	80	10	6996	334	99.94	9800+	58.60	29	13	7	626+	6
1	HC Brook (ENG)	2022-2023	12	20	1	1181	186	62.15	1287	91.76	4	7	1	141	23

In [4]:
```
# df.isnull().any()
```

In [5]:
```
df.isnull().sum()
```

Out[5]:
```
Player                0
Span                  0
Matches               0
Inns                  0
Not_Out               0
Runs                  0
Highest_in_score      0
Ave                   0
BF                    3
Batting_strike_rate   0
100                   0
50                    0
0                     0
4s                    0
6s                    0
dtype: int64
```

In [6]:
```
df[df['BF'].isnull()]
```

Out[6]:

	Player	Span	Matches	Inns	Not_Out	Runs	Highest_in_score	Ave	BF	Batting_strike_rate	100	50	0	4s	6s
8	ED Weekes (WI)	1948-1958	48	81	5	4455	207	58.61	NaN	0.0	15	19	6	258+	2
14	CL Walcott (WI)	1948-1960	44	74	7	3798	220	56.68	NaN	0.0	15	14	1	107+	11
55	Hon.FS Jackson (ENG)	1893-1905	20	33	4	1415	144*	48.79	NaN	0.0	5	6	3	51+	0

In [7]:
```
df['BF'] = df['BF'].fillna(0)
```

In [8]:
```
# checking null value is filled up or not
df[df['Player']== 'ED Weekes (WI)']
```

Out[8]:

	Player	Span	Matches	Inns	Not_Out	Runs	Highest_in_score	Ave	BF	Batting_strike_rate	100	50	0	4s	6s
8	ED Weekes (WI)	1948-1958	48	81	5	4455	207	58.61	0	0.0	15	19	6	258+	2

In [9]:
```
df[df['Player']== 'CL Walcott (WI)']
```

Out[9]:

	Player	Span	Matches	Inns	Not_Out	Runs	Highest_in_score	Ave	BF	Batting_strike_rate	100	50	0	4s	6s
14	CL Walcott (WI)	1948-1960	44	74	7	3798	220	56.68	0	0.0	15	14	1	107+	11

In [10]:
```
df[df['Player']== 'Hon.FS Jackson (ENG)']
```

Out[10]:

	Player	Span	Matches	Inns	Not_Out	Runs	Highest_in_score	Ave	BF	Batting_strike_rate	100	50	0	4s	6s
55	Hon.FS Jackson (ENG)	1893-1905	20	33	4	1415	144*	48.79	0	0.0	5	6	3	51+	0

In [11]:
```
# drop duplicate
df.duplicated()
```

Out[11]:
```
0     False
1     False
2     False
3     False
4     False
      ...
59     True
60    False
61    False
62    False
63    False
Length: 64, dtype: bool
```

In [12]:
```
df[df['Player'].duplicated()]
```

Out[12]:

	Player	Span	Matches	Inns	Not_Out	Runs	Highest_in_score	Ave	BF	Batting_strike_rate	100	50	0	4s	6s
51	V Kohli (IND)	2011-2024	113	191	11	8848	254*	49.15	15924	55.56	29	30	14	991	26
59	KD Walters (AUS)	1965-1981	74	125	14	5357	250	48.26	8662+	49.16	15	33	4	525+	23

In [13]:
```
df[df['Player'].isin(['V Kohli (IND)','KD Walters (AUS)'])]
```

Out[13]:

	Player	Span	Matches	Inns	Not_Out	Runs	Highest_in_score	Ave	BF	Batting_strike_rate	100	50	0	4s	6s
50	V Kohli (IND)	2011-2024	113	191	11	8848	254*	49.15	15924	55.56	29	30	14	991	26
51	V Kohli (IND)	2011-2024	113	191	11	8848	254*	49.15	15924	55.56	29	30	14	991	26
58	KD Walters (AUS)	1965-1981	74	125	14	5357	250	48.26	8662+	49.16	15	33	4	525+	23
59	KD Walters (AUS)	1965-1981	74	125	14	5357	250	48.26	8662+	49.16	15	33	4	525+	23

In [14]:
```
df = df.drop_duplicates()
```

In [15]:
```
# checking as the duplicates droped
df[df['Player'].isin(['V Kohli (IND)','KD Walters (AUS)'])]
```

Out[15]:

	Player	Span	Matches	Inns	Not_Out	Runs	Highest_in_score	Ave	BF	Batting_strike_rate	100	50	0	4s	6s
50	V Kohli (IND)	2011-2024	113	191	11	8848	254*	49.15	15924	55.56	29	30	14	991	26
58	KD Walters (AUS)	1965-1981	74	125	14	5357	250	48.26	8662+	49.16	15	33	4	525+	23

In [16]:
```
# split the span
df['Span'].str.split(pat = '-')
```

Out[16]:
```
0     [1928, 1948]
1     [2022, 2023]
2     [2015, 2016]
3     [1963, 1970]
4     [1930, 1954]
          ...
58    [1965, 1981]
60    [2002, 2014]
61    [1924, 1934]
62    [1930, 1938]
63    [1928, 1934]
Name: Span, Length: 62, dtype: object
```

In [17]:
```
df['First_Year'] = df['Span'].str.split(pat = '-').str[0]
```

In [18]:
```
df['Final_Year'] = df['Span'].str.split(pat = '-').str[1]
```

In [19]:
```
df.head(5)
```

Out[19]:

	Player	Span	Matches	Inns	Not_Out	Runs	Highest_in_score	Ave	BF	Batting_strike_rate	100	50	0	4s	6s
0	DG Bradman (AUS)	1928-1948	52	80	10	6996	334	99.94	9800+	58.60	29	13	7	626+	6
1	HC Brook (ENG)	2022-2023	12	20	1	1181	186	62.15	1287	91.76	4	7	1	141	23
2	AC Voges (AUS)	2015-2016	20	31	7	1485	269*	61.87	2667	55.68	5	4	2	186	5
3	RG Pollock (SA)	1963-1970	23	41	4	2256	274	60.97	1707+	54.48	7	11	1	246+	11
4	GA Headley (WI)	1930-1954	22	40	4	2190	270*	60.83	416+	56.00	10	5	2	104+	1

In [20]:
```
# not droping Span
df = df.drop(['Span'], axis = 1)
```

In [21]:
```
df.head(2)
```

Out[21]:

	Player	Matches	Inns	Not_Out	Runs	Highest_in_score	Ave	BF	Batting_strike_rate	100	50	0	4s	6s	First_Ye
0	DG Bradman (AUS)	52	80	10	6996	334	99.94	9800+	58.60	29	13	7	626+	6	192
1	HC Brook (ENG)	12	20	1	1181	186	62.15	1287	91.76	4	7	1	141	23	202

In [22]:
```
# split the country of the player

# Extract player's name and country
df['Player'].str.split(pat = '(')
```

Out[22]:
```
0      [DG Bradman , AUS)]
1        [HC Brook , ENG)]
2        [AC Voges , AUS)]
3       [RG Pollock , SA)]
4       [GA Headley , WI)]
              ...
58     [KD Walters , AUS)]
60    [GC Smith , ICC/SA)]
61    [WH Ponsford , AUS)]
62      [SJ McCabe , AUS)]
63     [DR Jardine , ENG)]
Name: Player, Length: 62, dtype: object
```

In [23]:
```
df['Country'] = df['Player'].str.split(pat = '(').str[1]
df['Country']
```

Out[23]:
```
0        AUS)
1        ENG)
2        AUS)
3         SA)
4         WI)
       ...
58       AUS)
60    ICC/SA)
61       AUS)
62       AUS)
63       ENG)
Name: Country, Length: 62, dtype: object
```

In [24]:
```
df['Country'] = df['Player'].str.split(pat = ')').str[0]
```

In [25]:
```
df['Country']
```

Out[25]:
```
0      DG Bradman (AUS
1        HC Brook (ENG
2        AC Voges (AUS
3       RG Pollock (SA
4       GA Headley (WI
            ...
58     KD Walters (AUS
60    GC Smith (ICC/SA
61    WH Ponsford (AUS
62      SJ McCabe (AUS
63     DR Jardine (ENG
Name: Country, Length: 62, dtype: object
```

In [26]:
```
df['Player'] = df['Player'].str.split(pat='(').str[0]
```

In [27]:
```
df.head()
```

Out[27]:

	Player	Matches	Inns	Not_Out	Runs	Highest_in_score	Ave	BF	Batting_strike_rate	100	50	0	4s	6s	First_Ye
0	DG Bradman	52	80	10	6996	334	99.94	9800+	58.60	29	13	7	626+	6	192
1	HC Brook	12	20	1	1181	186	62.15	1287	91.76	4	7	1	141	23	202
2	AC Voges	20	31	7	1485	269*	61.87	2667	55.68	5	4	2	186	5	201
3	RG Pollock	23	41	4	2256	274	60.97	1707+	54.48	7	11	1	246+	11	196
4	GA Headley	22	40	4	2190	270*	60.83	416+	56.00	10	5	2	104+	1	193

In [28]:
```
df['Highest_in_score'] = df['Highest_in_score'].str.split(pat='*').str[0]
df.head(5)
```

Out[28]:

	Player	Matches	Inns	Not_Out	Runs	Highest_in_score	Ave	BF	Batting_strike_rate	100	50	0	4s	6s	First_Ye
0	DG Bradman	52	80	10	6996	334	99.94	9800+	58.60	29	13	7	626+	6	192
1	HC Brook	12	20	1	1181	186	62.15	1287	91.76	4	7	1	141	23	202
2	AC Voges	20	31	7	1485	269	61.87	2667	55.68	5	4	2	186	5	201
3	RG Pollock	23	41	4	2256	274	60.97	1707+	54.48	7	11	1	246+	11	196
4	GA Headley	22	40	4	2190	270	60.83	416+	56.00	10	5	2	104+	1	193

In [29]:
```
df['Highest_in_score'].astype('int')
```

Out[29]:
```
0     334
1     186
2     269
3     274
4     270
     ...
58    250
60    277
61    266
62    232
63    127
Name: Highest_in_score, Length: 62, dtype: int32
```

In [30]:
```
df.dtypes
```

Out[30]:
```
Player                 object
Matches                 int64
Inns                    int64
Not_Out                 int64
Runs                    int64
Highest_in_score       object
Ave                   float64
BF                     object
Batting_strike_rate   float64
100                     int64
50                      int64
0                       int64
4s                     object
6s                     object
First_Year             object
Final_Year             object
Country                object
dtype: object
```

In [31]:
```
df=df.astype({'First_Year':'int','Final_Year':'int'})
```

In [32]:
```
df.dtypes
```

Out[32]:
```
Player                 object
Matches                 int64
Inns                    int64
Not_Out                 int64
Runs                    int64
Highest_in_score       object
Ave                   float64
BF                     object
Batting_strike_rate   float64
100                     int64
50                      int64
0                       int64
4s                     object
6s                     object
First_Year              int32
Final_Year              int32
Country                object
dtype: object
```

In [33]:
```
df['Matches'] = df['Matches'].astype('int')
```

In [34]:
```
df.dtypes
```

Out[34]:
```
Player                 object
Matches                 int32
Inns                    int64
Not_Out                 int64
Runs                    int64
Highest_in_score       object
Ave                   float64
BF                     object
Batting_strike_rate   float64
100                     int64
50                      int64
0                       int64
4s                     object
6s                     object
First_Year              int32
Final_Year              int32
Country                object
dtype: object
```

In [35]:
```
df['BF'] = df['BF'].str.split(pat='+').str[0]
```

In [36]:
```
df.head(2)
```

Out[36]:

	Player	Matches	Inns	Not_Out	Runs	Highest_in_score	Ave	BF	Batting_strike_rate	100	50	0	4s	6s	First_Year
0	DG Bradman	52	80	10	6996	334	99.94	9800	58.60	29	13	7	626+	6	1928
1	HC Brook	12	20	1	1181	186	62.15	1287	91.76	4	7	1	141	23	2022

In [37]:
```
df['4s'] = df['4s'].str.split(pat='+').str[0]
df
```

Out[37]:

	Player	Matches	Inns	Not_Out	Runs	Highest_in_score	Ave	BF	Batting_strike_rate	100	50	0	4s	6s	First_Y
0	DG Bradman	52	80	10	6996	334	99.94	9800	58.60	29	13	7	626	6	1
1	HC Brook	12	20	1	1181	186	62.15	1287	91.76	4	7	1	141	23	2
2	AC Voges	20	31	7	1485	269	61.87	2667	55.68	5	4	2	186	5	2
3	RG Pollock	23	41	4	2256	274	60.97	1707	54.48	7	11	1	246	11	1
4	GA Headley	22	40	4	2190	270	60.83	416	56.00	10	5	2	104	1	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
58	KD Walters	74	125	14	5357	250	48.26	8662	49.16	15	33	4	525	23	1
60	GC Smith	117	205	13	9265	277	48.25	15525	59.67	27	38	11	1165	24	2
61	WH Ponsford	29	48	4	2122	266	48.22	3118	44.77	7	6	1	119	0	1
62	SJ McCabe	39	62	5	2748	232	48.21	3217	60.02	6	13	4	241	5+	1
63	DR Jardine	22	33	6	1296	127	48.00	2110	25.59	1	10	2	53	0	1

62 rows × 17 columns