```
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
                                                                            BF
                                                                                  SR 100 50 0
                                                                                                    4s 6s
                                                                   Ave
                                                       6996
        0 DG Bradman (AUS) 1928-1948
                                                                  99.94 9800+
                                                                               58.60
                                                                                       29 13 7 626+
                                         52
                                              80
                                                   10
                                                             334
                                                                                                        6
         1
              HC Brook (ENG) 2022-2023
                                              20
                                                    1 1181
                                                             186
                                                                  62.15
                                                                          1287
                                                                               91.76
                                                                                                   141
        2
                                         20
                                              31
                                                    7 1485
                                                                  61.87
                                                                          2667 55.68
                                                                                                        5
              AC Voges (AUS) 2015-2016
                                                            269*
                                                                                              2
                                                                                                   186
                                                                                            4
         3
              RG Pollock (SA) 1963-1970
                                         23
                                              41
                                                      2256
                                                             274
                                                                  60.97
                                                                        1707+
                                                                                54.48
                                                                                        7 11 1 246+
                                         22
                                              40
                                                    4 2190 270* 60.83
                                                                         416+ 56.00
                                                                                            5 2 104+
         4
             GA Headley (WI) 1930-1954
                                                                                       10
In [2]: # rename col
        df = df.rename(columns = {'Mat': 'Matches','NO': 'Not_Out', 'HS': 'Highest_in_score', 'SR': 'Batting_strike_rate' })
       df.head(2)
In [3]:
Out[3]:
                                                                                   BF Batting_strike_rate 100 50 0
             Player
                     Span Matches Inns Not_Out Runs Highest_in_score
                                                                          Ave
                                                                                                                       4s 6s
                DG
                     1928-
                                52
                                      80
                                               10 6996
                                                                    334 99.94 9800+
         0 Bradman
                                                                                                  58.60
                                                                                                          29 13 7 626+
                                                                                                                           6
                     1948
              (AUS)
                HC
                     2022-
         1
                                 12
                                      20
                                                                    186 62.15
                                                                                                  91.76
                                                                                                                     141 23
              Brook
                                                1 1181
                                                                                 1287
                                                                                                              7 1
                     2023
              (ENG)
In [4]:
       # df.isnull().any()
In [5]: df.isnull().sum()
                                0
Out[5]:
        Player
         Span
                                0
        Matches
                                0
        Inns
                                0
                                0
        Not_Out
         Runs
        Highest_in_score
        Ave
                                0
         \mathsf{BF}
                                3
         Batting_strike_rate
                                0
         100
                                0
         50
                                0
         0
                                0
                                0
         4s
                                0
         6s
         dtype: int64
In [6]: df[df['BF'].isnull()]
Out[6]:
             Player Span Matches Inns Not_Out Runs Highest_in_score
                                                                                 BF Batting_strike_rate 100 50 0
                                                                          Ave
                                                                                                                     4s 6s
                ED
                    1948-
          8 Weekes
                                                                         58.61
                                      81
                                                5
                                                   4455
                                                                    207
                                                                                                   0.0
                                                                                                      15 19 6 258+
                                 48
                                                                               NaN
                      1958
               (WI)
                 CL
                     1948-
            Walcott
                                      74
                                                7 3798
                                                                    220 56.68 NaN
                                                                                                   0.0
                                                                                                        15 14 1 107+ 11
         14
                                 44
                     1960
               (WI)
             Hon.FS
                     1893-
                                                                                                         5 6 3
         55 Jackson
                                 20
                                      33
                                                4 1415
                                                                    144* 48.79 NaN
                                                                                                   0.0
                                                                                                                    51+
                                                                                                                          0
                     1905
              (ENG)
        df['BF'] = df['BF'].fillna(0)
In [7]:
In [8]: # checking null value is filled up or not
        df[df['Player']== 'ED Weekes (WI)']
Out[8]:
                      Span Matches Inns Not_Out Runs Highest_in_score
                                                                          Ave BF Batting_strike_rate 100 50 0
                                                                                                                     4s 6s
              Player
                 ED
                      1948-
         8
             Weekes
                                  48
                                       81
                                                 5 4455
                                                                      207 58.61
                                                                                  0
                                                                                                  0.0
                                                                                                       15 19 6 258+
                                                                                                                         2
                       1958
                (WI)
```

```
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
                  \mathsf{CL}
                       1948-
                                                                                                        15 14 1 107+ 11
              Walcott
                                   44
                                        74
                                                  7 3798
                                                                      220 56.68
                                                                                                   0.0
                        1960
                 (WI)
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
                Hon.FS
                        1893-
                                                   4 1415
          55
               Jackson
                                    20
                                         33
                                                                       144* 48.79
                                                                                   0
                                                                                                    0.0
                                                                                                          5
                                                                                                              6 3 51+
                                                                                                                          0
                         1905
                 (ENG)
In [11]: # drop duplicate
         df.duplicated()
Out[11]: 0
                False
          1
                False
          2
                False
          3
                False
          4
                False
                . . .
          59
                True
                False
          60
          61
                False
          62
                False
                False
          63
          Length: 64, dtype: bool
In [12]: df[df['Player'].duplicated()]
Out[12]:
                     Span Matches Inns Not_Out Runs Highest_in_score
                                                                          Ave
                                                                                   BF Batting_strike_rate 100 50
                                                                                                                        4s
                                                                                                                            6s
              V Kohli 2011-
                                113 191
                                                11 8848
                                                                    254* 49.15 15924
                                                                                                                       991 26
                                                                                                   55.56
                                                                                                          29
                                                                                                              30 14
               (IND)
                      2024
                 KD
                     1965-
            Walters
                                                                     250 48.26 8662+
                                                                                                          15 33
                                 74 125
                                                                                                   49.16
                                                                                                                 4 525+ 23
          59
                                                14 5357
                      1981
               (AUS)
In [13]: df[df['Player'].isin(['V Kohli (IND)','KD Walters (AUS)'])]
Out[13]:
                                                                                       Batting_strike_rate 100 50
              Player Span Matches Inns Not_Out Runs Highest_in_score
                                                                           Ave
                                                                                   BF
                                                                                                                        4s 6s
              V Kohli
                     2011-
          50
                                     191
                                                11 8848
                                                                    254* 49.15 15924
                                                                                                          29
                                113
                                                                                                   55.56
                                                                                                              30 14
                                                                                                                       991 26
                      2024
               (IND)
              V Kohli
                     2011-
          51
                                     191
                                                11 8848
                                                                    254* 49.15 15924
                                                                                                          29
                                                                                                              30 14
                                                                                                                       991 26
                      2024
               (IND)
                 KD
                     1965-
          58 Walters
                                 74 125
                                                14 5357
                                                                     250 48.26 8662+
                                                                                                   49.16
                                                                                                          15 33
                                                                                                                  4 525+ 23
                      1981
               (AUS)
                 KD
                      1965-
            Walters
                                 74 125
                                                14 5357
                                                                                                                 4 525+ 23
                                                                     250 48.26 8662+
                                                                                                   49.16
                                                                                                          15 33
                      1981
               (AUS)
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]:
                                                                                   BF Batting strike rate 100 50
              Player Span Matches Inns Not_Out Runs Highest_in_score
                                                                           Ave
                                                                                                                            6s
              V Kohli
                     2011-
          50
                                     191
                                                11 8848
                                                                    254* 49.15 15924
                                                                                                   55.56
                                                                                                          29
                                                                                                              30 14
                                                                                                                       991
                                                                                                                           26
                                113
               (IND)
                      2024
                 KD
                     1965-
          58 Walters
                                                                     250 48.26 8662+
                                                                                                   49.16
                                                                                                          15 33
                                                                                                                  4 525+
                                 74 125
                                                14 5357
                                                                                                                            23
                      1981
               (AUS)
In [16]: # split the span
         df['Span'].str.split(pat = '-')
```

```
Out[16]: 0
                [1928, 1948]
          1
                [2022, 2023]
                [2015, 2016]
          2
          3
                [1963, 1970]
                [1930, 1954]
          4
                [1965, 1981]
          58
                [2002, 2014]
          60
          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
                 DG
                      1928-
          0 Bradman
                                  52
                                       80
                                                 10
                                                   6996
                                                                      334 99.94 9800+
                                                                                                     58.60
                                                                                                            29
                                                                                                               13 7 626+
                                                                                                                              6
                       1948
               (AUS)
                 HC
                      2022-
          1
                                  12
                                       20
                                                  1 1181
                                                                      186 62.15
                                                                                   1287
                                                                                                     91.76
                                                                                                                 7 1
                                                                                                                        141 23
               Brook
                       2023
               (ENG)
                 AC
                      2015-
                                  20
                                       31
                                                  7 1485
                                                                      269* 61.87
                                                                                   2667
                                                                                                     55.68
                                                                                                                   2
                                                                                                                        186
                                                                                                                              5
               Voges
                       2016
               (AUS)
                  RG
                      1963-
          3
              Pollock
                                  23
                                       41
                                                  4 2256
                                                                      274 60.97 1707+
                                                                                                     54.48
                                                                                                             7 11 1 246+ 11
                       1970
                 (SA)
                 GΑ
                      1930-
             Headley
                                  22
                                       40
                                                  4 2190
                                                                      270* 60.83
                                                                                  416+
                                                                                                     56.00
                                                                                                            10
                                                                                                                 5 2 104+
                       1954
                 (WI)
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
                                                                              BF Batting_strike_rate 100 50 0
                                                                                                                   4s 6s First_Year
                                                                      Ave
                 DG
          0 Bradman
                           52
                                 80
                                          10 6996
                                                                334 99.94 9800+
                                                                                              58.60
                                                                                                      29 13 7 626+
                                                                                                                               192
               (AUS)
                 HC
                                                                                                                               202
         1
                           12
                                 20
                                           1 1181
                                                                186 62.15
                                                                            1287
                                                                                              91.76
                                                                                                       4 7 1
                                                                                                                141 23
               Brook
               (ENG)
In [22]: # split the country of the player
         # Extract player's name and country
          df['Player'].str.split(pat = '(')
Out[22]:
                 [DG Bradman , AUS)]
         0
          1
                   [HC Brook , ENG)]
                   [AC Voges , AUS)]
          3
                  [RG Pollock , SA)]
                  [GA Headley , WI)]
          58
                 [KD Walters , AUS)]
                [GC Smith , ICC/SA)]
          60
          61
                [WH Ponsford , AUS)]
                  [SJ McCabe , AUS)]
          62
                 [DR Jardine , ENG)]
          63
          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]:
                 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
                 DR Jardine (ENG
          63
          Name: Country, Length: 62, dtype: object
In [26]: df['Player'] = df['Player'].str.split(pat='(').str[0]
In [27]:
         df.head()
Out[27]:
                                                                                                                  4s 6s First_Year
              Player Matches Inns Not_Out Runs Highest_in_score
                                                                     Ave
                                                                             BF Batting_strike_rate 100 50 0
                 DG
                                                               334 99.94 9800+
          0
                                80
                                          10 6996
                                                                                                     29 13 7 626+
                                                                                                                              192
                           52
                                                                                             58.60
            Bradman
                 HC
                                                                                                                              202
                                20
                                                                                             91.76
         1
                           12
                                           1 1181
                                                               186 62.15
                                                                            1287
                                                                                                        7 1
                                                                                                                 141 23
               Brook
                 AC
         2
                           20
                                31
                                                              269* 61.87
                                                                                             55.68
                                                                                                         4 2
                                                                                                                 186
                                                                                                                     5
                                                                                                                              201
                                           7 1485
                                                                            2667
               Voges
                 RG
         3
                                                                                                                              196
                           23
                                41
                                           4 2256
                                                               274 60.97 1707+
                                                                                             54.48
                                                                                                      7 11 1 246+ 11
              Pollock
                 GΑ
                                                                                                                              193
                           22
                                40
                                           4 2190
                                                              270* 60.83
                                                                           416+
                                                                                             56.00
                                                                                                     10
                                                                                                         5 2 104+
             Headley
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_Year
                           52
                                80
                                          10 6996
                                                               334 99.94 9800+
                                                                                             58.60
                                                                                                    29 13 7 626+
                                                                                                                              192
            Bradman
                 HC
         1
                                                                                                               141 23
                           12
                                20
                                           1 1181
                                                               186 62.15
                                                                           1287
                                                                                             91.76
                                                                                                         7 1
                                                                                                                              202
               Brook
                 AC
         2
                           20
                                31
                                           7 1485
                                                               269 61.87
                                                                            2667
                                                                                             55.68
                                                                                                         4 2
                                                                                                                 186
                                                                                                                     5
                                                                                                                              201
               Voges
                 RG
         3
                           23
                                41
                                           4 2256
                                                               274 60.97 1707+
                                                                                             54.48
                                                                                                      7 11 1 246+ 11
                                                                                                                              196
              Pollock
                 GΑ
                           22
                                40
                                           4 2190
                                                               270 60.83
                                                                           416+
                                                                                             56.00
                                                                                                     10
                                                                                                         5 2 104+
                                                                                                                              193
             Headley
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]:
                                   object
          Player
                                    int64
          Matches
                                    int64
          Inns
                                    int64
          Not_Out
          Runs
                                    int64
                                   object
          Highest_in_score
                                  float64
          Ave
          \mathsf{BF}
                                   object
                                  float64
          Batting_strike_rate
          100
                                    int64
          50
                                    int64
          0
                                    int64
          4s
                                   object
          6s
                                   object
          First_Year
                                   object
          Final_Year
                                   object
          Country
                                   object
          dtype: object
          df=df.astype({'First_Year':'int','Final_Year':'int'})
In [31]:
          df.dtypes
In [32]:
                                   object
Out[32]: Player
          Matches
                                    int64
          Inns
                                    int64
          Not_Out
                                    int64
          Runs
                                    int64
                                   object
          Highest_in_score
                                  float64
          Ave
          \mathsf{BF}
                                   object
          Batting_strike_rate
                                  float64
          100
                                    int64
          50
                                    int64
          0
                                    int64
          4s
                                   object
                                   object
          6s
          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
                                    int64
          Inns
                                    int64
          Not_Out
          Runs
                                    int64
          Highest_in_score
                                   object
          Ave
                                  float64
          \mathsf{BF}
                                   object
          Batting_strike_rate
                                  float64
          100
                                    int64
          50
                                    int64
          0
                                    int64
          4s
                                   object
                                   object
          6s
          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 E	Batting_strike_rate	100 !	50 C		4s 6s	Fii	rst_Year
	0	DG Bradman	52	80	10	6996	334	99.94	9800	58.60	29	13 7	62	6+ 6	5	1928
	1	HC Brook	12	20	1	1181	186	62.15	1287	91.76	4	7 1	1	41 23	3	2022
	4															•
In [37]:	df[	['4s'] = c	lf['4s'].s	str.sp	lit(pat='	+').st	r[0]									
Out[37]:		Player	Matches	Inns	Not_Out	Runs	Highest_in_score	Ave	BF	Batting_strike_rate	100	50	0	4s	6s	First_\
	0	DG Bradman	52	80	10	6996	334	99.94	9800	58.60	29	13	7	626	6	1
	1	HC Brook	1)	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
	•••						<b></b>	•••	•••							
	58	KD Walters	//	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		62	5	2748	232	48.21	3217	60.02	6	13	4	241	5+	1
	63	DR Jardine		33	6	1296	127	48.00	2110	25.59	1	10	2	53	0	1
	62 r	ows × 17 c	columns													