PYTHON PROGRAMMING

ENDSEM PROJECT

IPL DATA ANALYSIS

NAME : Dheeraj
ENROLL NO. : 200330

ABOUT THE FILE
'IPL Ball-by-Ball 2008-2020.csv' - Contains the ball by ball data of the whole Indian Premier League tournament from year 2008 to 2020
'IPL Matches 2008-2020.csv' - Contains the Information of Matches like teams , venue , umpires etc. , of the whole Indian Premier League tournament from year 2008 to 2020

Data Link : https://www.kaggle.com/patrickb1912/ipl-complete-dataset-20082020

In [1]:
```
#importing libraries and files
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings("ignore")
B_Data = pd.read_csv("IPL Ball-by-Ball 2008-2020.csv")
M_Data = pd.read_csv("IPL Matches 2008-2020.csv")
```

Beginning with the basic Steps , since we have two csv files , we can create a class 'CSVFileInfo' and make the two files as objects of this class and then simple call the 'display_summary' method to get all the basic info of the file.

In [2]:
```
class CSVFileInfo:
    def __init__(self, path, file_name):
        self.path = path
        self.file_name = file_name
    def display_summary(self):
        try:
            data = pd.read_csv(self.path + self.file_name)
        except FileNotFoundError:
            print("File Not Found , check file path again !")
        else:
            print(self.file_name)
            print(data.info())
            print(data.head())
            print(data.tail())
            print(data.isnull().sum())
```

In [3]:
```
BallData = CSVFileInfo("C:\\Users\\Dheeraj Mehlawat\\Desktop\\4th Semester\\Python Programming\\ENDSEM PROJECT\\", "
MatchData = CSVFileInfo("C:\\Users\\Dheeraj Mehlawat\\Desktop\\4th Semester\\Python Programming\\ENDSEM PROJECT\\",
```

In [4]:
```
newdata = CSVFileInfo("C:\\Users\\Dheeraj Mehlawat\\Deskt","NewData")
```

In [5]:
```
newdata.display_summary()
```

```
File Not Found , check file path again !
```

In [6]:
```
1 BallData.display_summary()
```

```
IPL Ball-by-Ball 2008-2020.csv
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 193468 entries, 0 to 193467
Data columns (total 18 columns):
 #   Column            Non-Null Count   Dtype
---  ------            --------------   -----
 0   id                193468 non-null  int64
 1   inning            193468 non-null  int64
 2   over              193468 non-null  int64
 3   ball              193468 non-null  int64
 4   batsman           193468 non-null  object
 5   non_striker       193468 non-null  object
 6   bowler            193468 non-null  object
 7   batsman_runs      193468 non-null  int64
 8   extra_runs        193468 non-null  int64
 9   total_runs        193468 non-null  int64
 10  non_boundary      193468 non-null  int64
 11  is_wicket         193468 non-null  int64
 12  dismissal_kind    9495 non-null    object
 13  player_dismissed  9495 non-null    object
 14  fielder           6784 non-null    object
 15  extras_type       10233 non-null   object
 16  batting_team      193468 non-null  object
 17  bowling_team      193468 non-null  object
dtypes: int64(9), object(9)
memory usage: 26.6+ MB
None
       id  inning  over  ball      batsman  non_striker     bowler  \
0  335982       1     6     5   RT Ponting  BB McCullum  AA Noffke
1  335982       1     6     6  BB McCullum   RT Ponting  AA Noffke
2  335982       1     7     1  BB McCullum   RT Ponting     Z Khan
3  335982       1     7     2  BB McCullum   RT Ponting     Z Khan
4  335982       1     7     3   RT Ponting  BB McCullum     Z Khan

   batsman_runs  extra_runs  total_runs  non_boundary  is_wicket  \
0             1           0           1             0          0
1             1           0           1             0          0
2             0           0           0             0          0
3             1           0           1             0          0
```

```
4             1           0           1             0          0

  dismissal_kind player_dismissed fielder extras_type           batting_team  \
0            NaN              NaN     NaN         NaN  Kolkata Knight Riders
1            NaN              NaN     NaN         NaN  Kolkata Knight Riders
2            NaN              NaN     NaN         NaN  Kolkata Knight Riders
3            NaN              NaN     NaN         NaN  Kolkata Knight Riders
4            NaN              NaN     NaN         NaN  Kolkata Knight Riders

                 bowling_team
0  Royal Challengers Bangalore
1  Royal Challengers Bangalore
2  Royal Challengers Bangalore
3  Royal Challengers Bangalore
4  Royal Challengers Bangalore
            id  inning  over  ball  batsman non_striker           bowler  \
193463  1237181       1    12     5  RR Pant     SS Iyer  NM Coulter-Nile
193464  1237181       1    12     6  RR Pant     SS Iyer  NM Coulter-Nile
193465  1237181       1    13     1  RR Pant     SS Iyer        KH Pandya
193466  1237181       1    13     2  RR Pant     SS Iyer        KH Pandya
193467  1237181       1    13     3  SS Iyer     RR Pant        KH Pandya

        batsman_runs  extra_runs  total_runs  non_boundary  is_wicket  \
193463             0           0           0             0          0
193464             1           0           1             0          0
193465             0           1           1             0          0
193466             1           0           1             0          0
193467             1           0           1             0          0

       dismissal_kind player_dismissed fielder extras_type    batting_team  \
193463            NaN              NaN     NaN         NaN  Delhi Capitals
193464            NaN              NaN     NaN         NaN  Delhi Capitals
193465            NaN              NaN     NaN       wides  Delhi Capitals
193466            NaN              NaN     NaN         NaN  Delhi Capitals
193467            NaN              NaN     NaN         NaN  Delhi Capitals

          bowling_team
193463  Mumbai Indians
193464  Mumbai Indians
193465  Mumbai Indians
193466  Mumbai Indians
193467  Mumbai Indians
```

```
id                      0
inning                  0
over                    0
ball                    0
batsman                 0
non_striker             0
bowler                  0
batsman_runs            0
extra_runs              0
total_runs              0
non_boundary            0
is_wicket               0
dismissal_kind     183973
player_dismissed   183973
fielder            186684
extras_type        183235
batting_team            0
bowling_team            0
dtype: int64
```

We have null values in 4 columns that is 'dismissal_kind' , 'player_dismissed' , 'fielder' and 'extras_type' , which is justified as these columns can only have values if a player is being dismissed , got caught or runout by a feilder , or if extras are given on a delivery . so this data is valid to work upon and no rows need to be dropped or removed .

In [7]:
```
1 MatchData.display_summary()
```

```
IPL Matches 2008-2020.csv
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 816 entries, 0 to 815
Data columns (total 17 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   id               816 non-null    int64
 1   city             803 non-null    object
 2   date             816 non-null    object
 3   player_of_match  812 non-null    object
 4   venue            816 non-null    object
 5   neutral_venue    816 non-null    int64
 6   team1            816 non-null    object
 7   team2            816 non-null    object
 8   toss_winner      816 non-null    object
 9   toss_decision    816 non-null    object
 10  winner           812 non-null    object
 11  result           812 non-null    object
 12  result_margin    799 non-null    float64
 13  eliminator       812 non-null    object
 14  method           19 non-null     object
 15  umpire1          816 non-null    object
 16  umpire2          816 non-null    object
dtypes: float64(1), int64(2), object(14)
memory usage: 108.5+ KB
None
       id        city        date player_of_match  \
0  335982   Bangalore  2008-04-18     BB McCullum
1  335983  Chandigarh  2008-04-19      MEK Hussey
2  335984       Delhi  2008-04-19     MF Maharoof
3  335985      Mumbai  2008-04-20      MV Boucher
4  335986     Kolkata  2008-04-20       DJ Hussey

                                      venue  neutral_venue  \
0                     M Chinnaswamy Stadium              0
1  Punjab Cricket Association Stadium, Mohali              0
2                          Feroz Shah Kotla              0
3                          Wankhede Stadium              0
4                              Eden Gardens              0
```

```
                         team1                         team2  \
0  Royal Challengers Bangalore        Kolkata Knight Riders
1              Kings XI Punjab          Chennai Super Kings
2             Delhi Daredevils             Rajasthan Royals
3               Mumbai Indians  Royal Challengers Bangalore
4        Kolkata Knight Riders              Deccan Chargers

                   toss_winner toss_decision                        winner  \
0  Royal Challengers Bangalore         field        Kolkata Knight Riders
1          Chennai Super Kings           bat          Chennai Super Kings
2             Rajasthan Royals           bat             Delhi Daredevils
3               Mumbai Indians           bat  Royal Challengers Bangalore
4              Deccan Chargers           bat        Kolkata Knight Riders

    result  result_margin eliminator method    umpire1         umpire2
0     runs          140.0          N    NaN  Asad Rauf     RE Koertzen
1     runs           33.0          N    NaN  MR Benson      SL Shastri
2  wickets            9.0          N    NaN  Aleem Dar  GA Pratapkumar
3  wickets            5.0          N    NaN   SJ Davis       DJ Harper
4  wickets            5.0          N    NaN  BF Bowden     K Hariharan
          id       city        date player_of_match  \
811  1216547      Dubai  2020-09-28  AB de Villiers
812  1237177      Dubai  2020-11-05       JJ Bumrah
813  1237178  Abu Dhabi  2020-11-06   KS Williamson
814  1237180  Abu Dhabi  2020-11-08      MP Stoinis
815  1237181      Dubai  2020-11-10        TA Boult

                                  venue  neutral_venue  \
811  Dubai International Cricket Stadium              0
812  Dubai International Cricket Stadium              0
813                Sheikh Zayed Stadium              0
814                Sheikh Zayed Stadium              0
815  Dubai International Cricket Stadium              0

                         team1                team2          toss_winner  \
811  Royal Challengers Bangalore       Mumbai Indians       Mumbai Indians
812              Mumbai Indians       Delhi Capitals       Delhi Capitals
813  Royal Challengers Bangalore  Sunrisers Hyderabad  Sunrisers Hyderabad
814              Delhi Capitals  Sunrisers Hyderabad       Delhi Capitals
815              Delhi Capitals       Mumbai Indians       Delhi Capitals
```

```
    toss_decision                     winner   result  result_margin  \
811         field  Royal Challengers Bangalore      tie            NaN
812         field               Mumbai Indians     runs           57.0
813         field          Sunrisers Hyderabad  wickets            6.0
814           bat               Delhi Capitals     runs           17.0
815           bat               Mumbai Indians  wickets            5.0

    eliminator method      umpire1      umpire2
811          Y    NaN  Nitin Menon   PR Reiffel
812          N    NaN  CB Gaffaney  Nitin Menon
813          N    NaN   PR Reiffel       S Ravi
814          N    NaN   PR Reiffel       S Ravi
815          N    NaN  CB Gaffaney  Nitin Menon
id                 0
city              13
date               0
player_of_match    4
venue              0
neutral_venue      0
team1              0
team2              0
toss_winner        0
toss_decision      0
winner             4
result             4
result_margin     17
eliminator         4
method           797
umpire1            0
umpire2            0
dtype: int64
```

1. BEST PLAYERS IN ALL CATEGORIES

1.1 TOP RUN SCORER

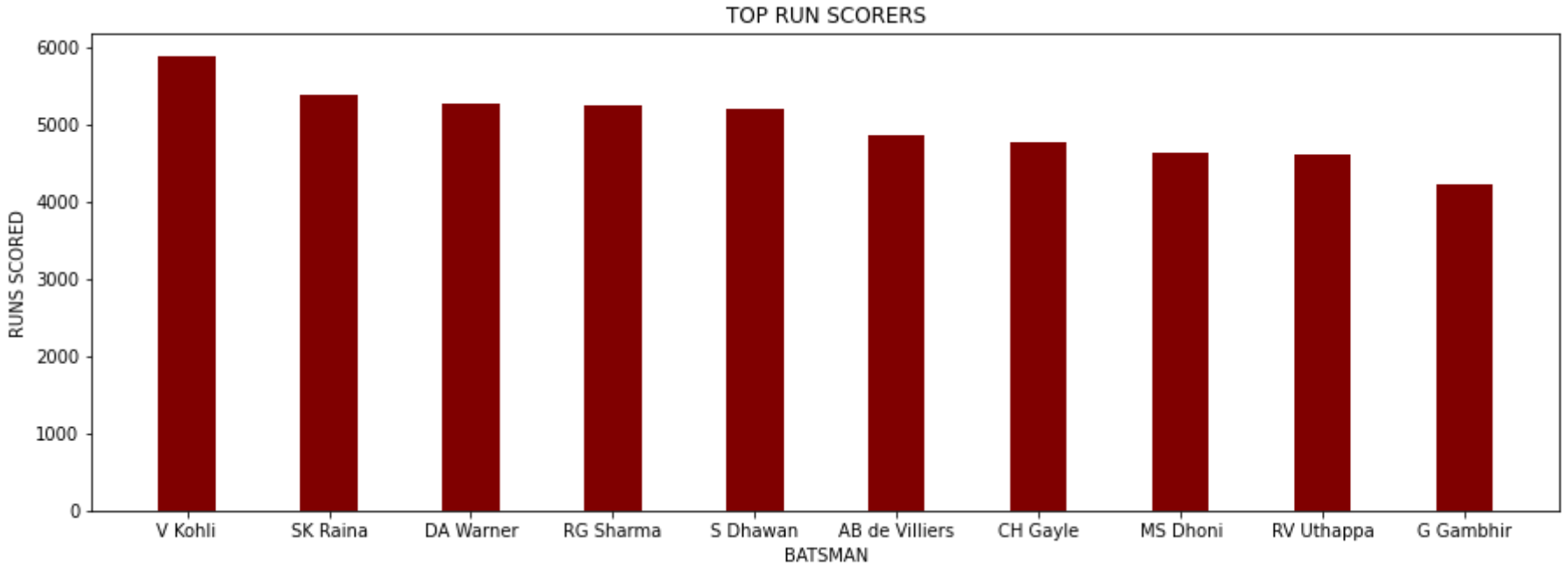
In [8]:
```
1 run_scorers = B_Data.groupby('batsman')['batsman_runs'].sum().sort_values(ascending = False).head(10).reset_index().
```

In [9]:
```
1 run_scorers
```

Out[9]:

	batsman	Total_Runs
0	V Kohli	5878
1	SK Raina	5368
2	DA Warner	5254
3	RG Sharma	5230
4	S Dhawan	5197
5	AB de Villiers	4849
6	CH Gayle	4772
7	MS Dhoni	4632
8	RV Uthappa	4607
9	G Gambhir	4217

In [10]:
```
fig = plt.figure(figsize = (15, 5))
plt.bar(run_scorers['batsman'],run_scorers['Total_Runs'],color ='maroon',width = 0.4)
plt.title("TOP RUN SCORERS")
plt.xlabel("BATSMAN")
plt.ylabel("RUNS SCORED")
plt.show()
```



1.2 TOP WICKET TAKER

In [11]:
```
1 B_Data['dismissal_kind'].unique()
```

Out[11]:
```
array([nan, 'caught', 'run out', 'bowled', 'lbw', 'retired hurt',
       'stumped', 'caught and bowled', 'hit wicket',
       'obstructing the field'], dtype=object)
```

Among these only those wickets are considered as bowler's wicket where dismissal kind is 'caught' , 'bowled' , 'lbw', 'stumped' , 'caught and bowled' or 'hit wicket'

In [12]:
```
1 BowlersWicket = [ 'caught' , 'bowled' , 'lbw', 'stumped' , 'caught and bowled' , 'hit wicket']
```

In [13]:
```
1 BowlerWicketData = B_Data[B_Data['dismissal_kind'].isin(BowlersWicket)]
```

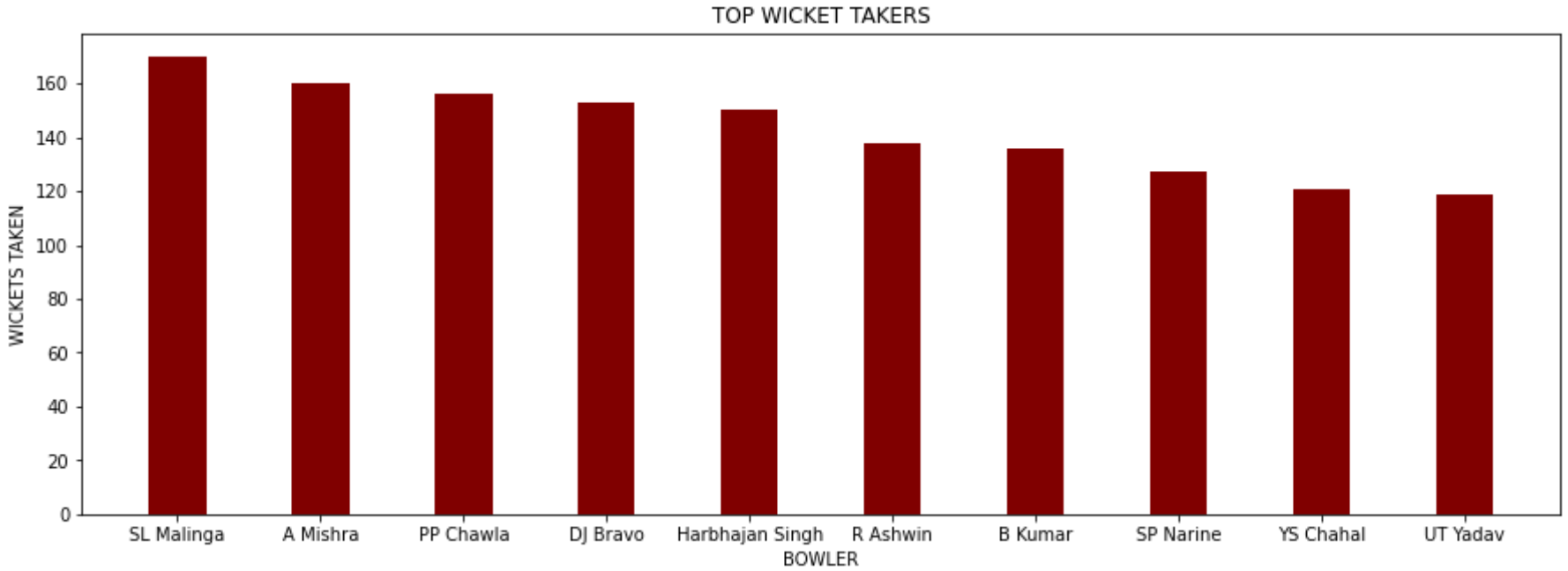
In [14]:
```
1 wicket_takers = BowlerWicketData.groupby('bowler')['is_wicket'].count().sort_values(ascending = False).head(10).rese
```

In [15]:
```
1 wicket_takers
```

Out[15]:

	bowler	Total_Wickets
0	SL Malinga	170
1	A Mishra	160
2	PP Chawla	156
3	DJ Bravo	153
4	Harbhajan Singh	150
5	R Ashwin	138
6	B Kumar	136
7	SP Narine	127
8	YS Chahal	121
9	UT Yadav	119

In [16]:
```
fig = plt.figure(figsize = (15, 5))
plt.bar(wicket_takers['bowler'],wicket_takers['Total_Wickets'],color ='maroon',width = 0.4)
plt.title("TOP WICKET TAKERS")
plt.xlabel("BOWLER")
plt.ylabel("WICKETS TAKEN")
plt.show()
```



1.3 BATSMEN WITH MOST BOUNDARIES (4's + 6's)

In [17]:
```
1 BoundaryData = B_Data[B_Data['batsman_runs'].isin([4,6])].reset_index()
```

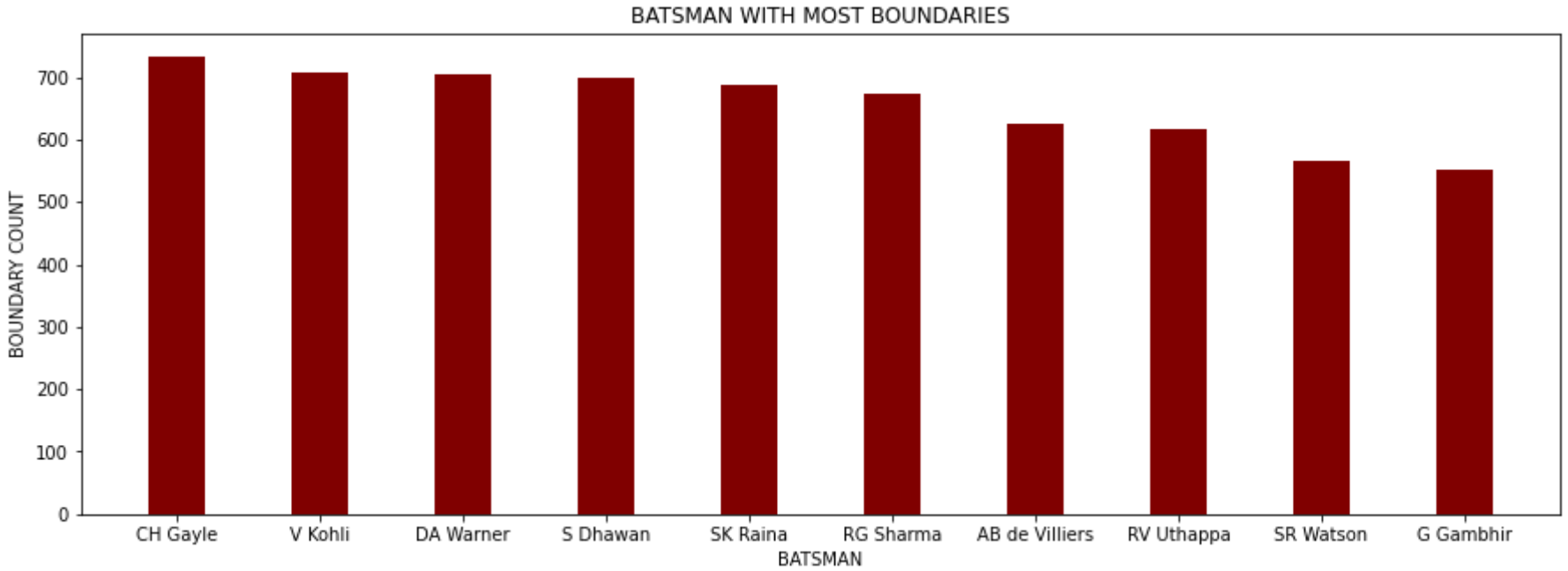
In [18]:
```
1 boundaries = BoundaryData.groupby('batsman')['batsman_runs'].count().sort_values(ascending = False).head(10).reset_i
```

In [19]:
```
1 boundaries
```

Out[19]:

	batsman	Total_Boundaries
0	CH Gayle	733
1	V Kohli	706
2	DA Warner	705
3	S Dhawan	700
4	SK Raina	687
5	RG Sharma	672
6	AB de Villiers	625
7	RV Uthappa	617
8	SR Watson	566
9	G Gambhir	551

In [20]:
```
fig = plt.figure(figsize = (15, 5))
plt.bar(boundaries['batsman'],boundaries['Total_Boundaries'],color ='maroon',width = 0.4)
plt.title("BATSMAN WITH MOST BOUNDARIES")
plt.xlabel("BATSMAN")
plt.ylabel("BOUNDARY COUNT")
plt.show()
```



1.4 MOST 'PLAYER OF THE MATCH' AWARDS

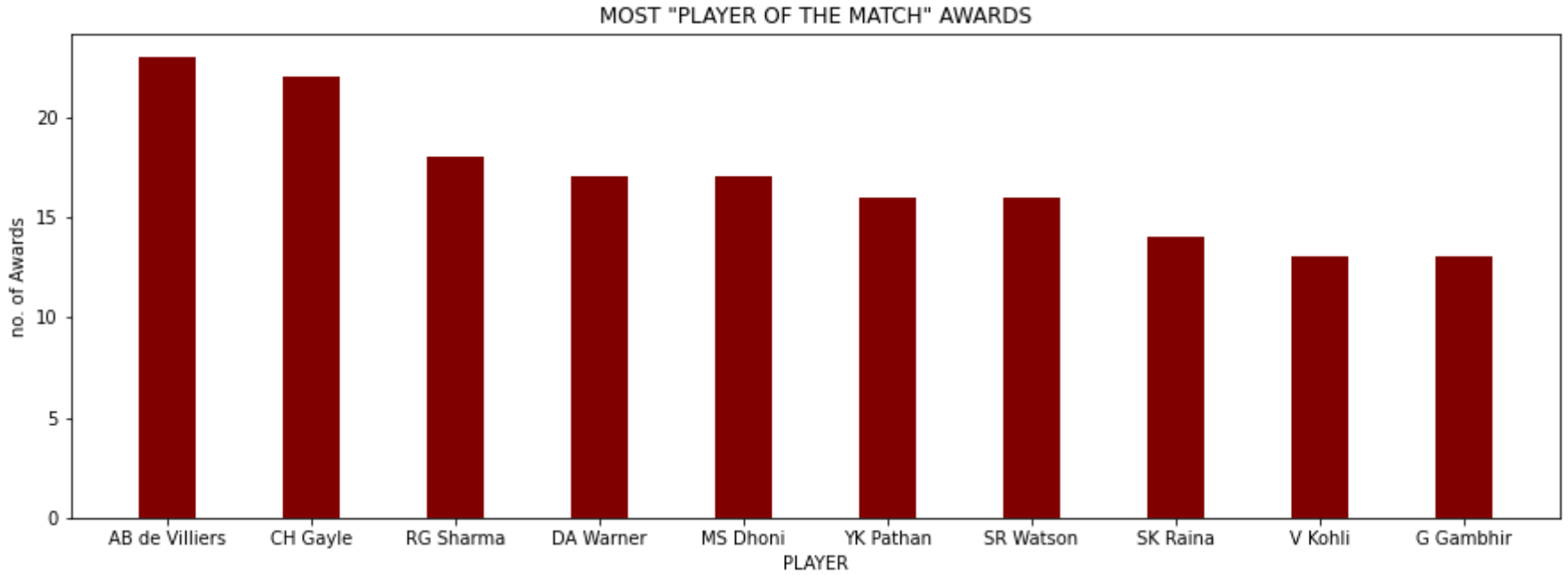
In [21]:
```
playerofmatch = M_Data.groupby('player_of_match')['id'].count().sort_values(ascending = False).head(10).reset_index(
```

In [22]:
```
1 playerofmatch
```

Out[22]:

	Player Name	Award count
0	AB de Villiers	23
1	CH Gayle	22
2	RG Sharma	18
3	DA Warner	17
4	MS Dhoni	17
5	YK Pathan	16
6	SR Watson	16
7	SK Raina	14
8	V Kohli	13
9	G Gambhir	13

In [23]:
```
fig = plt.figure(figsize = (15, 5))
plt.bar(playerofmatch['Player Name'],playerofmatch['Award count'],color ='maroon',width = 0.4)
plt.title('MOST "PLAYER OF THE MATCH" AWARDS')
plt.xlabel("PLAYER")
plt.ylabel("no. of Awards")
plt.show()
```



1.5 MOST 50+ SCORES BY A BATSMAN

In [24]:
```
PlayerScores = B_Data.groupby(['id','batsman'])['batsman_runs'].sum().reset_index().rename(columns={'id':'50+ Scores
```

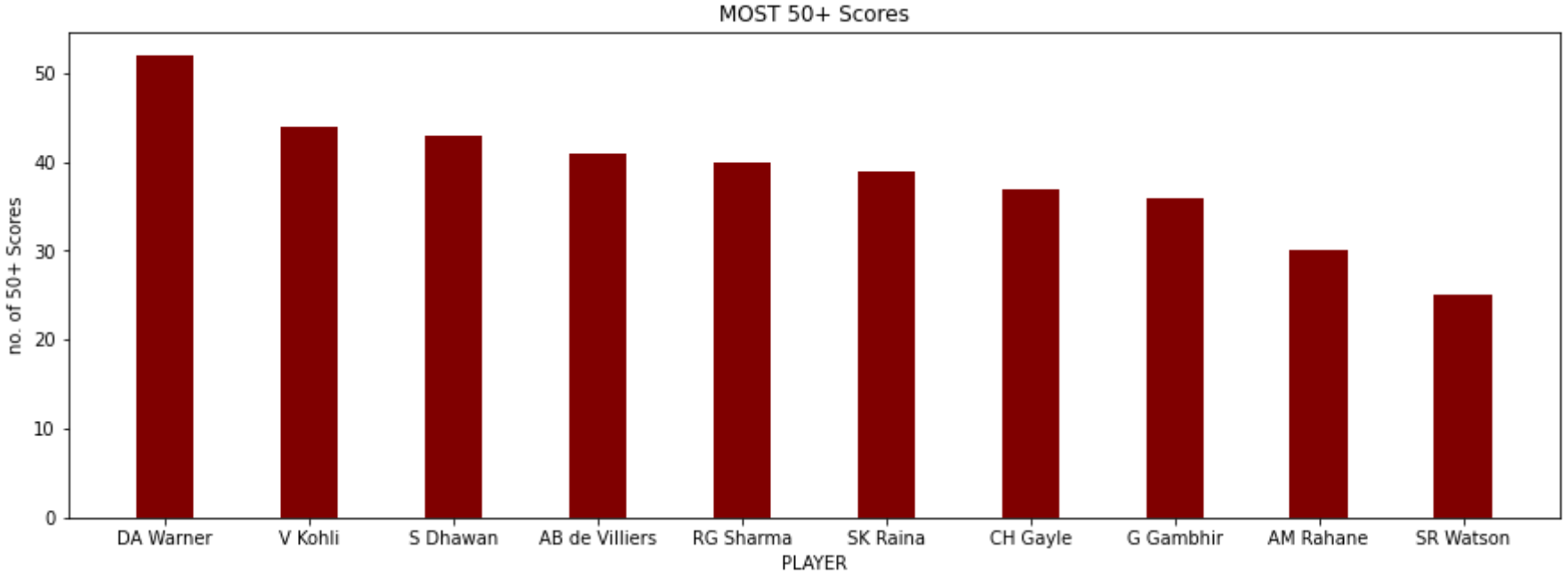
In [25]:
```
fiftyplus = PlayerScores[PlayerScores['batsman_runs']>=50].groupby(['batsman']).count().sort_values(by = 'batsman_ru
```

In [26]:
```
1 fiftyplus
```

Out[26]:

	batsman	50+ Scores
0	DA Warner	52
1	V Kohli	44
2	S Dhawan	43
3	AB de Villiers	41
4	RG Sharma	40
5	SK Raina	39
6	CH Gayle	37
7	G Gambhir	36
8	AM Rahane	30
9	SR Watson	25

In [27]:
```
fig = plt.figure(figsize = (15, 5))
plt.bar(fiftyplus['batsman'],fiftyplus['50+ Scores'],color ='maroon',width = 0.4)
plt.title('MOST 50+ Scores')
plt.xlabel("PLAYER")
plt.ylabel("no. of 50+ Scores")
plt.show()
```



2. TOP RIVALRIES

2.1 BOWLERS DISMISSING A PARTICULAR BATSMAN MOST TIMES

In [28]:
```
1 BowlerData = BowlerWicketData[['bowler','is_wicket','dismissal_kind','player_dismissed']]
```

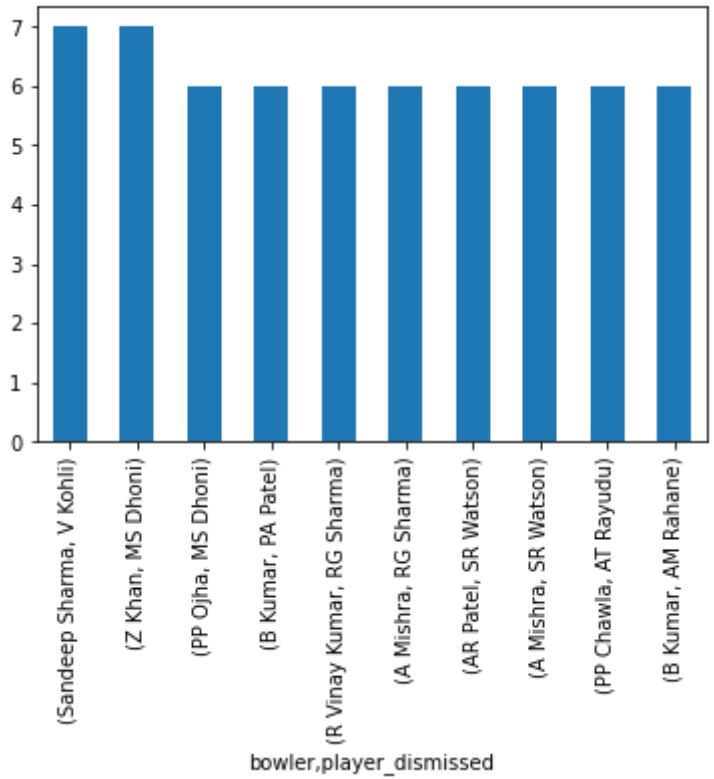
In [29]:
```
1 bowlervsbatsman = BowlerData.groupby(['bowler','player_dismissed'])['is_wicket'].sum().sort_values(ascending = False
```

In [30]:
```
1 bowlervsbatsman
```

Out[30]:
```
bowler          player_dismissed
Sandeep Sharma  V Kohli            7
Z Khan          MS Dhoni           7
PP Ojha         MS Dhoni           6
B Kumar         PA Patel           6
R Vinay Kumar   RG Sharma          6
A Mishra        RG Sharma          6
AR Patel        SR Watson          6
A Mishra        SR Watson          6
PP Chawla       AT Rayudu          6
B Kumar         AM Rahane          6
Name: is_wicket, dtype: int64
```

In [31]:
```
1 bowlervsbatsman.plot(x='bowler',y='No.of wickets',kind='bar')
```

Out[31]: <AxesSubplot:xlabel='bowler,player_dismissed'>



2.2 BATSMEN SCORING MOST RUNS AGAINST A PARTICULAR BOWLER

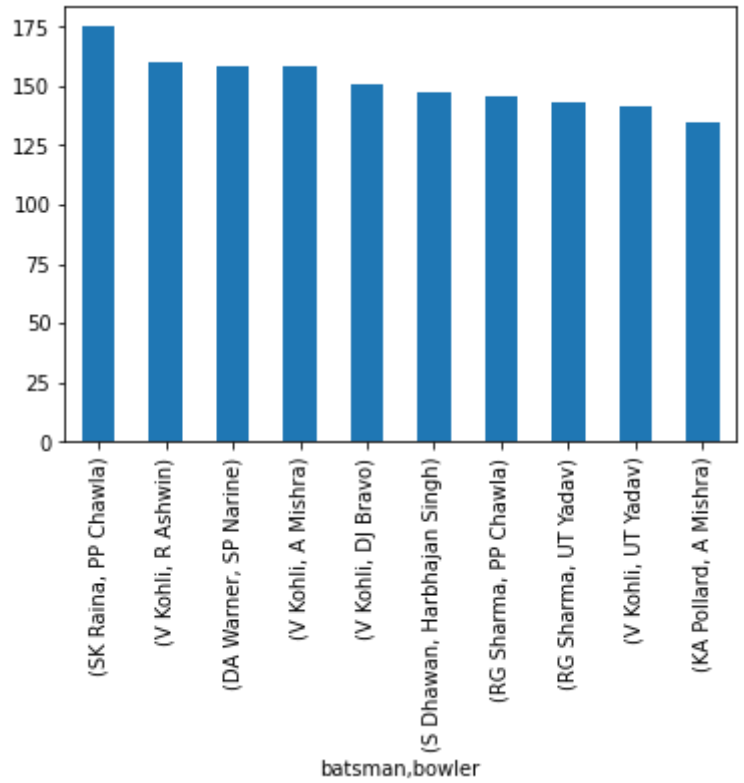
In [32]:
```
batsmanvsbowler = B_Data.groupby(['batsman','bowler'])['batsman_runs'].sum().sort_values(ascending = False).head(10)
```

In [33]:
```
batsmanvsbowler
```

Out[33]:
```
batsman     bowler
SK Raina    PP Chawla          175
V Kohli     R Ashwin           160
DA Warner   SP Narine          158
V Kohli     A Mishra           158
            DJ Bravo           151
S Dhawan    Harbhajan Singh    147
RG Sharma   PP Chawla          146
            UT Yadav           143
V Kohli     UT Yadav           141
KA Pollard  A Mishra           135
Name: batsman_runs, dtype: int64
```

In [34]:
```
1 batsmanvsbowler.plot(x='batsman',y='batsman_runs',kind='bar')
```

Out[34]: <AxesSubplot:xlabel='batsman,bowler'>



3. SOME BOWLERS STATS

3.1 BOWLERS WHO GAVE MOST EXTRAS

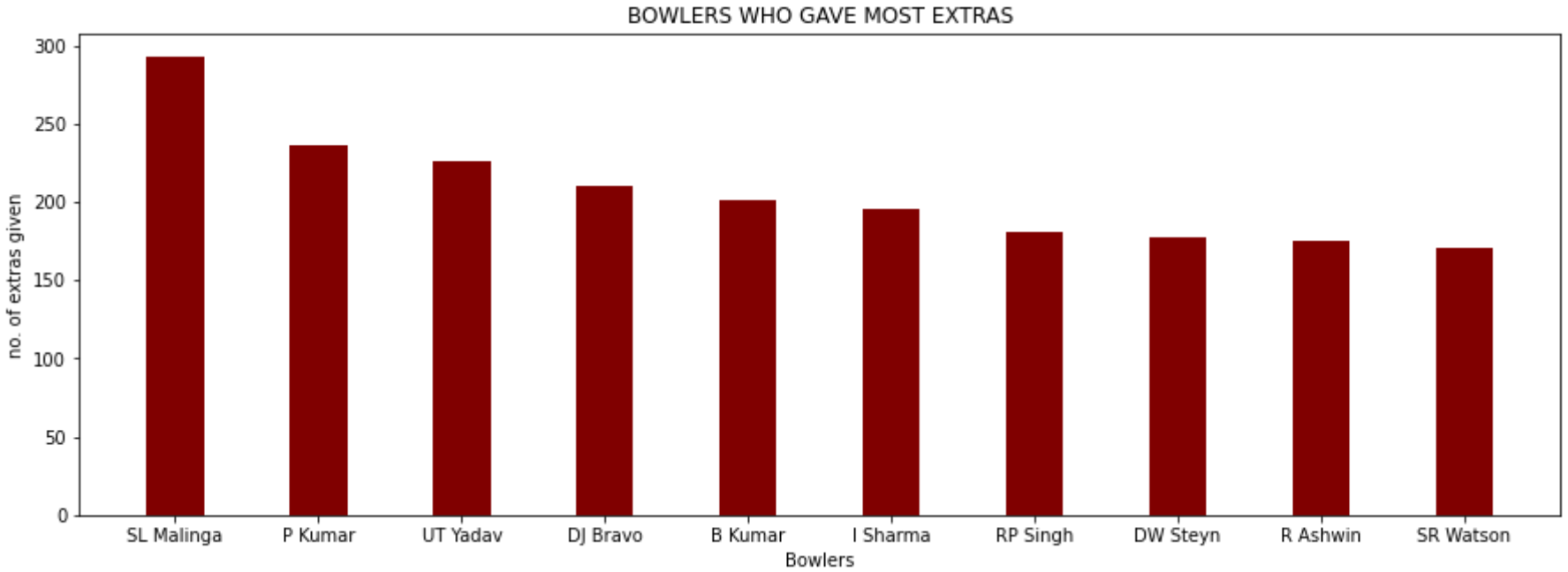
In [35]:
```
extraruns = B_Data.groupby('bowler')['extra_runs'].sum().sort_values(ascending = False).head(10).reset_index()
```

In [36]:
```
extraruns
```

Out[36]:

	bowler	extra_runs
0	SL Malinga	293
1	P Kumar	236
2	UT Yadav	226
3	DJ Bravo	210
4	B Kumar	201
5	I Sharma	196
6	RP Singh	181
7	DW Steyn	177
8	R Ashwin	175
9	SR Watson	171

In [37]:
```
fig = plt.figure(figsize = (15, 5))
plt.bar(extraruns['bowler'],extraruns['extra_runs'],color ='maroon',width = 0.4)
plt.title('BOWLERS WHO GAVE MOST EXTRAS')
plt.xlabel("Bowlers")
plt.ylabel("no. of extras given")
plt.show()
```



3.2 BOWLERS CONCEIVING MOST BOUNDARIES

In [38]:
```
BoundaryData = B_Data[B_Data['batsman_runs'].isin([4,6])].reset_index()
```

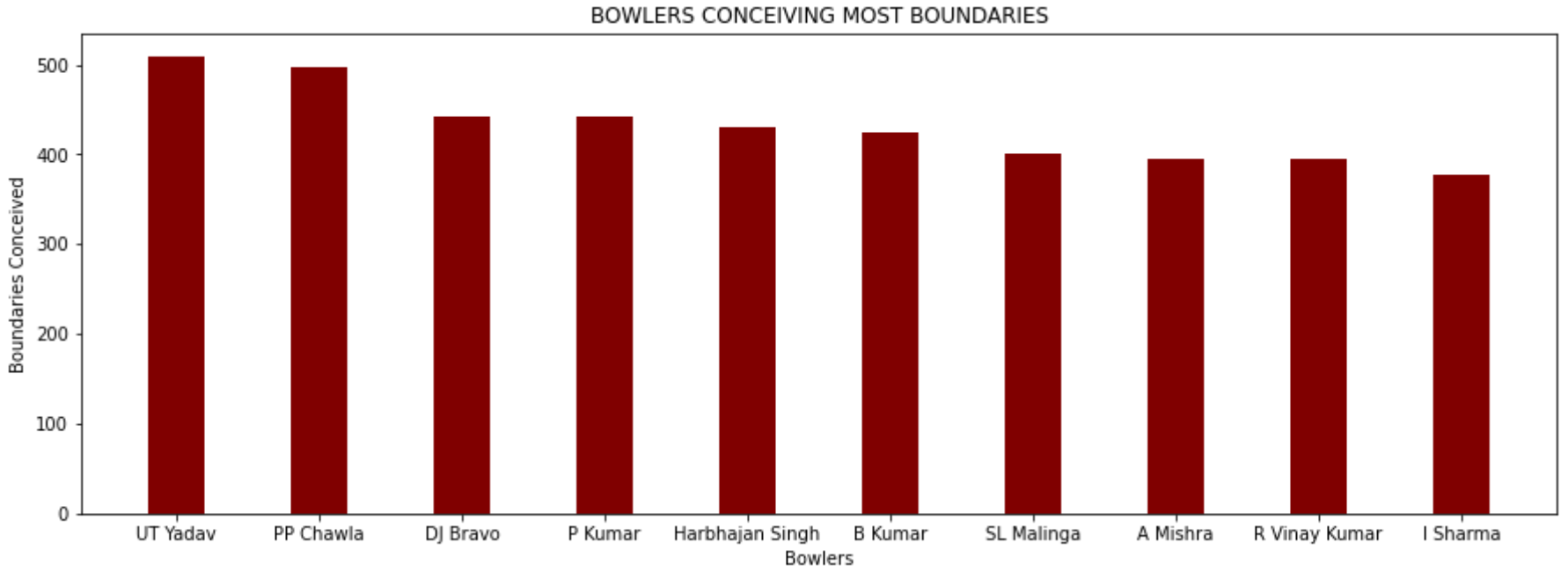
In [39]:
```
mostboundaries = BoundaryData.groupby('bowler')['batsman_runs'].count().sort_values(ascending = False).head(10).rese
```

In [40]:
```
1 mostboundaries
```

Out[40]:

	bowler	Boundaries Conceived
0	UT Yadav	509
1	PP Chawla	497
2	DJ Bravo	443
3	P Kumar	442
4	Harbhajan Singh	431
5	B Kumar	424
6	SL Malinga	400
7	A Mishra	394
8	R Vinay Kumar	394
9	I Sharma	378

In [41]:
```
fig = plt.figure(figsize = (15, 5))
plt.bar(mostboundaries['bowler'],mostboundaries['Boundaries Conceived'],color ='maroon',width = 0.4)
plt.title('BOWLERS CONCEIVING MOST BOUNDARIES')
plt.xlabel("Bowlers")
plt.ylabel("Boundaries Conceived")
plt.show()
```



4. OTHER STATS

4.1 TOP 5 VENUES WITH MOST MATCHES

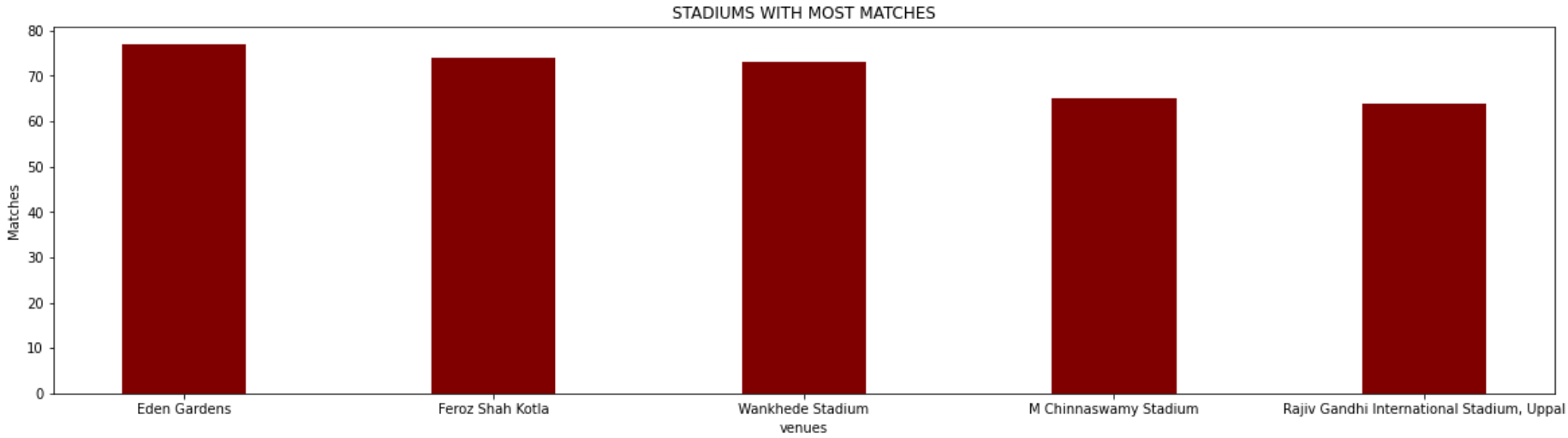
In [42]:
```
venuedata = M_Data.groupby('venue')['id'].count().sort_values(ascending = False).head(5).reset_index().rename(column
```

In [43]:
```
venuedata
```

Out[43]:

	venue	Matches
0	Eden Gardens	77
1	Feroz Shah Kotla	74
2	Wankhede Stadium	73
3	M Chinnaswamy Stadium	65
4	Rajiv Gandhi International Stadium, Uppal	64

In [51]:
```
fig = plt.figure(figsize = (20, 5))
plt.bar(venuedata['venue'],venuedata['Matches'],color ='maroon',width = 0.4)
plt.title('STADIUMS WITH MOST MATCHES')
plt.xlabel("venues")
plt.ylabel("Matches")
plt.show()
```



4.2 BATTING FIRST VS BATTING SECOND

In [45]:
```
1 Batting_first = M_Data[M_Data['result']=='runs']['id'].count()
```

In [46]:
```
1 Batting_first
```

Out[46]: 364

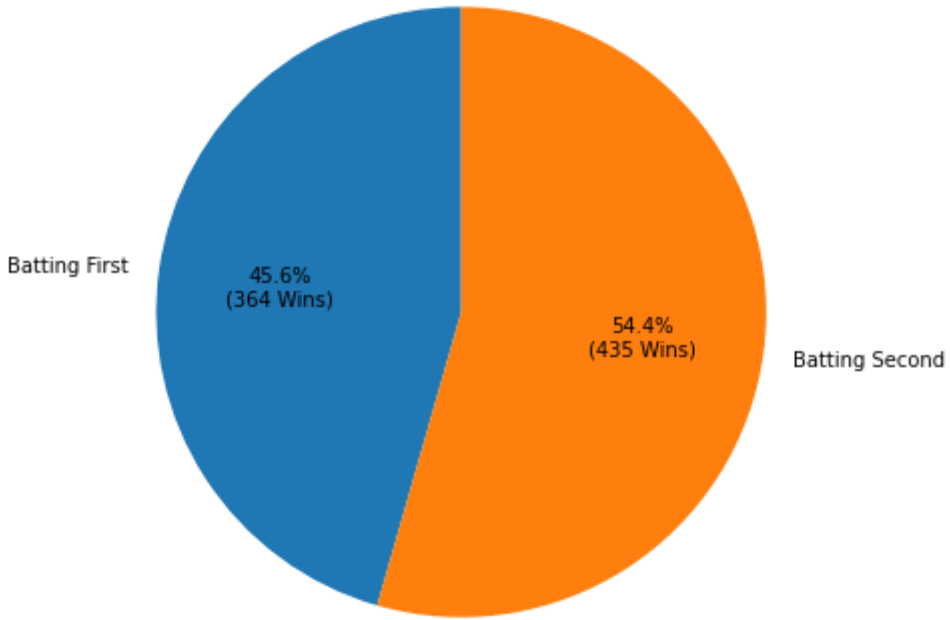
In [47]:
```
1 Batting_second = M_Data[M_Data['result']=='wickets']['id'].count()
```

In [48]:
```
1 Batting_second
```

Out[48]: 435

In [49]:
```
def func(pct, allvalues):
    absolute = int(pct / 100.*np.sum(allvalues))
    return "{:.1f}%\n({:d} Wins)".format(pct, absolute)
y =[364,435]
mylabels = ["Batting First", "Batting Second"]
fig, ax = plt.subplots(figsize =(10, 7))
ax.pie(y,autopct = lambda pct: func(pct, y),labels = mylabels,startangle = 90)

plt.show()
```



4.3 Five Most Expensive Overs in IPL history

In [50]:
```
1 B_Data.groupby(['id','inning','over','bowler'])['batsman_runs'].sum().sort_values(ascending = False).head(5).reset_i
```

Out[50]:

	id	inning	over	bowler	batsman_runs
0	501247	2	2	P Parameswaran	36
1	734047	2	5	P Awana	32
2	548327	2	12	R Sharma	31
3	335988	2	12	A Symonds	30
4	980987	1	18	S Kaushik	30

In [ ]:
```
1
```