ANUSHKA PRIYA
RA2011003010555

Importing Libraries

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
sns.set_style("darkgrid")
from warnings import filterwarnings
filterwarnings('ignore')
```

Reading Dataset:

```
match_data = pd.read_csv("../input/ipl-complete-dataset-20082020/IPL Matches 2008-2020.csv")
ball_data = pd.read_csv("../input/ipl-complete-dataset-20082020/IPL Ball-by-Ball 2008-2020.csv")

match_data.head()
```

```
       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
```

```
ball_data.head()
```

```
       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
```

```
match_data.info()
```

```
<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
```

```
match_data.describe()
```

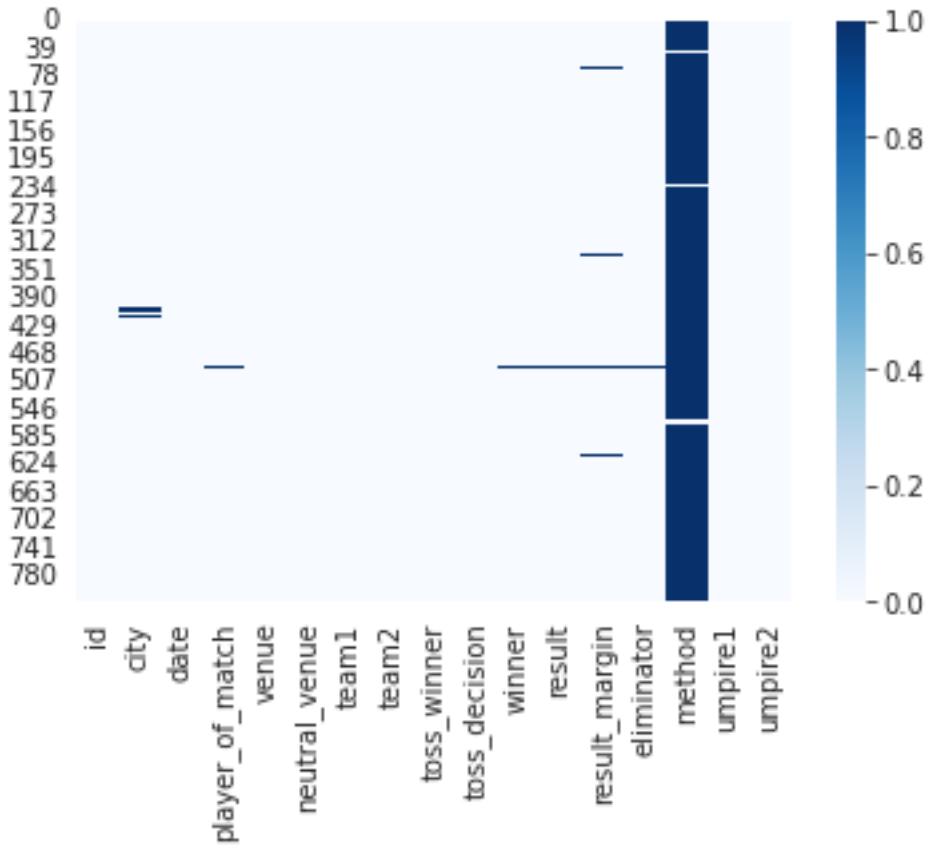
```
                 id  neutral_venue  result_margin
count  8.160000e+02     816.000000     799.000000
mean   7.563496e+05       0.094363      17.321652
std    3.058943e+05       0.292512      22.068427
min    3.359820e+05       0.000000       1.000000
25%    5.012278e+05       0.000000       6.000000
50%    7.292980e+06       0.000000       8.000000
75%    1.082626e+06       0.000000      19.500000
max    1.237181e+06       1.000000     146.000000
```

```
match_data.dtypes.value_counts()
```

```
object     14
int64       2
float64     1
dtype: int64
```

Checking for Null values:

```
sns.heatmap(match_data.isnull(),cmap='Blues')
plt.show()
```




```
match_data.isnull().sum()
```

```
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
```

```
print("The percentage of null values in the columns consisting of null
values:\n")
for i in match_data.columns:
```

```
    null_rate = match_data[i].isna().sum() / len(match_data) * 100
    if null_rate > 0 :
        print("null rate of {} column: {}
%".format(i,round(null_rate,3)))
```

```
The percentage of null values in the columns consisting of null
values:

null rate of city column: 1.593%
null rate of player_of_match column: 0.49%
null rate of winner column: 0.49%
null rate of result column: 0.49%
null rate of result_margin column: 2.083%
null rate of eliminator column: 0.49%
null rate of method column: 97.672%
```

Number of unique values in each column:

```
match_data.nunique()
```

```
id                816
city               32
date              596
player_of_match   233
venue              36
neutral_venue       2
team1              15
team2              15
toss_winner        15
toss_decision       2
winner             15
result              3
result_margin      91
eliminator          2
method              1
umpire1            48
umpire2            47
dtype: int64
```

```
match_data.columns
```

```
Index(['id', 'city', 'date', 'player_of_match', 'venue',
'neutral_venue',
       'team1', 'team2', 'toss_winner', 'toss_decision', 'winner',
'result',
       'result_margin', 'eliminator', 'method', 'umpire1', 'umpire2'],
      dtype='object')
```

```
match_data.shape
```

```
(816, 17)
```

Data preprocessing:

City:

```
match_data[match_data['city'].isnull()]
```

```
        id city        date player_of_match  \
399  729281  NaN  2014-04-17       YS Chahal
402  729287  NaN  2014-04-19        PA Patel
403  729289  NaN  2014-04-19       JP Duminy
404  729291  NaN  2014-04-20      GJ Maxwell
406  729295  NaN  2014-04-22      GJ Maxwell
407  729297  NaN  2014-04-23       RA Jadeja
408  729299  NaN  2014-04-24         CA Lynn
409  729301  NaN  2014-04-25        AJ Finch
410  729303  NaN  2014-04-25       MM Sharma
413  729309  NaN  2014-04-27         M Vijay
414  729311  NaN  2014-04-27        DR Smith
415  729313  NaN  2014-04-28  Sandeep Sharma
417  729317  NaN  2014-04-30         B Kumar

                                    venue  neutral_venue  \
399              Sharjah Cricket Stadium              1
402  Dubai International Cricket Stadium              1
403  Dubai International Cricket Stadium              1
404              Sharjah Cricket Stadium              1
406              Sharjah Cricket Stadium              1
407  Dubai International Cricket Stadium              1
408              Sharjah Cricket Stadium              1
409  Dubai International Cricket Stadium              1
410  Dubai International Cricket Stadium              1
413              Sharjah Cricket Stadium              1
414              Sharjah Cricket Stadium              1
415  Dubai International Cricket Stadium              1
417  Dubai International Cricket Stadium              1

                          team1                         team2  \
399             Delhi Daredevils  Royal Challengers Bangalore
402  Royal Challengers Bangalore              Mumbai Indians
403       Kolkata Knight Riders            Delhi Daredevils
404            Rajasthan Royals             Kings XI Punjab
406             Kings XI Punjab         Sunrisers Hyderabad
407            Rajasthan Royals         Chennai Super Kings
408  Royal Challengers Bangalore       Kolkata Knight Riders
409         Sunrisers Hyderabad            Delhi Daredevils
410         Chennai Super Kings              Mumbai Indians
413            Delhi Daredevils              Mumbai Indians
414         Sunrisers Hyderabad          Chennai Super Kings
415             Kings XI Punjab  Royal Challengers Bangalore
417              Mumbai Indians          Sunrisers Hyderabad
```

```
                    toss_winner toss_decision
winner  \
399  Royal Challengers Bangalore         field  Royal Challengers
Bangalore
402  Royal Challengers Bangalore         field  Royal Challengers
Bangalore
403        Kolkata Knight Riders           bat             Delhi
Daredevils
404               Kings XI Punjab         field              Kings XI
Punjab
406          Sunrisers Hyderabad         field              Kings XI
Punjab
407             Rajasthan Royals         field          Chennai Super
Kings
408  Royal Challengers Bangalore         field        Kolkata Knight
Riders
409          Sunrisers Hyderabad           bat          Sunrisers
Hyderabad
410               Mumbai Indians           bat          Chennai Super
Kings
413               Mumbai Indians           bat             Delhi
Daredevils
414          Sunrisers Hyderabad           bat          Chennai Super
Kings
415              Kings XI Punjab         field              Kings XI
Punjab
417               Mumbai Indians         field          Sunrisers
Hyderabad

      result  result_margin eliminator method          umpire1
umpire2
399  wickets            8.0          N    NaN        Aleem Dar
S Ravi
402  wickets            7.0          N    NaN        Aleem Dar    AK
Chaudhary
403  wickets            4.0          N    NaN        Aleem Dar     VA
Kulkarni
404  wickets            7.0          N    NaN        BF Bowden       M
Erasmus
406     runs           72.0          N    NaN        M Erasmus
S Ravi
407     runs            7.0          N    NaN  HDPK Dharmasena  RK
Illingworth
408     runs            2.0          N    NaN        Aleem Dar     VA
Kulkarni
409     runs            4.0          N    NaN        M Erasmus
S Ravi
410  wickets            7.0          N    NaN        BF Bowden       M
Erasmus
413  wickets            6.0          N    NaN        Aleem Dar     VA
```

```
Kulkarni
414  wickets            5.0          N    NaN     AK Chaudhary     VA
Kulkarni
415  wickets            5.0          N    NaN        BF Bowden
S Ravi
417     runs           15.0          N    NaN  HDPK Dharmasena       M
Erasmus
```

From the given table consisting of **null value present in "city" column** we are able to infer that

- Rows consisting of **Sharjah or Dubai stadium** only have their "city" values missing

Hence we would be filling the null values accordingly

```
for idx in match_data[match_data['city'].isna()].index:
        match_data.loc[idx, 'city'] = 'Sharjah' if match_data.loc[idx,
'venue'] == 'Sharjah Cricket Stadium' else 'Dubai'
```

#Let us take an look at the unique values present in the city column

```
match_data['city'].unique()
```

```
array(['Bangalore', 'Chandigarh', 'Delhi', 'Mumbai', 'Kolkata',
'Jaipur',
       'Hyderabad', 'Chennai', 'Cape Town', 'Port Elizabeth',
'Durban',
       'Centurion', 'East London', 'Johannesburg', 'Kimberley',
       'Bloemfontein', 'Ahmedabad', 'Cuttack', 'Nagpur', 'Dharamsala',
       'Kochi', 'Indore', 'Visakhapatnam', 'Pune', 'Raipur', 'Ranchi',
       'Abu Dhabi', 'Sharjah', 'Dubai', 'Rajkot', 'Kanpur',
'Bengaluru'],
      dtype=object)
```

We get to see **Bangalore** as well as **Bengaluru** , hence we will replace it appropriately

```
match_data['city'].replace('Bengaluru', 'Bangalore', inplace=True)
```