**IPL**

**SUPER**

**PREDICTOR**

**(IBM HACK CHALLENGE - 2021**)



Done by:

Abhinandan Kumar,

Mano Ranjith. R,

Monisha. D,

Vigneshwaran. K

**CONTENTS**

|  |  |
| --- | --- |
| **S.NO** | **TITLE** |
| 1. | Introduction |
| 2. | Data Processing |
| 3. | IBM Cognos Analysis |
| 4. | Python Notebook |
| 5. | Results |
| 6. | Important Links |



**INTRODUCTION**

A lot of thing in this world is unpredictable and one of them which keeps us at the edge of seat and has an unpredictable outcome is IPL – Indian Premiere League. It is a series of cricket tournament conducted between various teams for 50+ days. But everything in life can be predictable with a degree of precision if the data is available. Let us try with the given dataset to form insights of various matches.

The project chosen by us is “**IPL SUPER PREDICTOR**”. The problem statement is as discussed below.

**Description:**  
Since the dawn of the IPL in 2008, it has attracted viewers all around the globe. A high level of uncertainty and last moment nail biters has drawn the fans to watch the matches in large numbers. Within a short period, IPL has become the highest revenue-generating league of cricket. With all this, the amount of data being generated in terms of matches revenue scores, etc has also become huge. Analysing such vast amounts of data would give great insights in forecasting match results, top scores, and wicket-takers, etc.

**Expected Solution:**

The objective of this solution is to create a dashboard that visualizes the following capabilities and also forecast the future results

1. To find the team that won the greatest number of matches in the entire IPL.
2. To find the team that lost the greatest number of matches in the entire IPL.
3. Does winning a toss increase the chances of victory.
4. To find the player with the most player of the match awards.
5. To find the city that hosted the maximum number of IPL matches.
6. To find the most winning team for each season.
7. To find the on-field umpire with the maximum number of IPL matches.
8. To find the biggest victories in IPL while defending a total and while chasing a total.
9. Which team won the most matches while batting first
10. Which team won the most matches while batting second
11. List of teams which have won matches by most runs cumulatively

Hence the various analysis of data is done by using technologies mentioned below.

**TECHNOLOGY STACK:**

IBM Tools: IBM CLOUD, IBM COGNOS ANALYTICS

Language: PYTHON

Python Libraries: NumPy, Pandas, matplotlib, Seaborn

Implemented in Jupyter Notebook and IBM Dashboard.

**DATA PROCESSING**

The dataset was downloaded from the given link and it was carefully looked through to find various relationships and identifying columns, their importance pertaining to problem statements and their necessity. After such analysis the following steps were done to process data and make them ready for data analysis.

**a) Dropping unnecessary columns**

After looking at the dataset few unnecessary columns were identified. These columns were removed from the dataset to keep it clean. These columns were - *id, date, dl\_applied, venue, umpire3.*

**b) Handling Synonymous Values**

There were two values that meant the same but were written differently. They were:

* Bengaluru and Bangalore
* S Ravi and Sundaram Ravi

The Bengaluru was changed to Bangalore and Sundaram Ravi to S Ravi. These were done to avoid splitting data of total count under separate name that mean the same.

**c) Handling Missing Values**

There were few values missing under the columns of City, Umpire1, Umpire2 and Man of the Match columns. The values of these matches were searched and identified from ESPN site and inserted into the dataset. The missing values were identified by plotting a *heatmap* as show in the notebook.

**NOTE:** Few matches were cancelled thus winner outcome cannot be identified for such matches. To handle such values a proxy value called ‘Over’ was inserted to the dataset.

**d) Adding additional required columns**

New columns were created and appended to the end of dataset as and when required for special addressal of specific problem statements. Columns added were – Loser, Wining Both Toss and Match, Batting First and Batting Second.

**COGNOS ANALYTICS**

The prepared data consisted of matchesCopy.csv, Deliveries.csv, Sheet2.csv and Problem6.csv. These were uploaded to the IBM Cloud to be used for analysis.

Steps to be Followed:

1) Launching IBM Cognos Analytics

2) Uploading the Dataset to the IBM Cloud.

3) Graphs that were plotted using the given dataset were:

* Using the Deliveries.csv the total runs, wickets, matches and overs were calculated and the **cards** were added to the dashboard. An extra
* **Bar Chart** was plotted for the following – Most Matches Won, Most Matches Lost, Teams that won both Match and Toss, Batting First and Second Count
* **Packed Bubble** Graph was plotted for - On Field Umpires (Umpire 1 and Umpire 2), Player of the Matches, Cities count
* A regression line was drawn for TossWins vs Wins and points were plotted for TossWins vs Loss.
* For cities, **a geographical mapping** was done along with the Packed Bubble Graph already mentioned.
* A **Table** was added to the dashboard showing most winning by season count.
* A **Line Plot** was plotted for biggest winning by most runs cumulatively.
* Used a function called count() to count the values for host of match locations ,total winning teams etc.

4) As the final step the Entire Dashboard was exported as PDF and the link to the dashboard was saved.

**JUPYTER NOTEBOOK**

The analysis of data was done using Python and its libraries like NumPy, Pandas, matplotlib, Seaborn. The libraries were first imported and necessary changes were done to its data. At last, the changes done were served in place by turning *inplace = True*. These changed Data Frame was exported to a new .csv file.

Following Steps were Performed on the Data:

**1) Analysis and Processing of Data:**

Unnecessary columns were dropped, missing values were identified and added to the dataset, Synonymous values were unified, Additional required columns were appended to the dataset.

**2) Plotting and Analysing Each Problem Statement:**

Various graphs were plotted and each and every statement were analysed to draw necessary conclusions. Graphs included – Scatter Plot, Bar Graph, Regression Lines, Pie Chart etc. Sperate Data Frames were also created and addressed.

At times when necessary separated functions were created and using apply () function necessary data transformations were done.

At the end precise answers were drawn from the analysis.

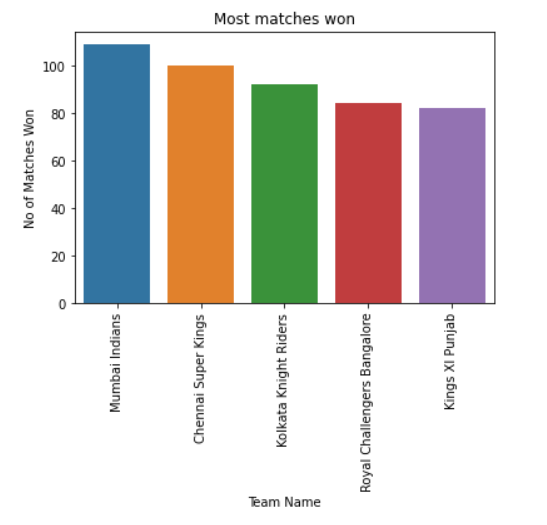
**3) Conclusion:**

At the end after all necessary analysis, the questions asked in problem statement were successfully answered and a conclusion was drawn. It is presented at the end of notebook for reference as summary.

This processed data was used to confirm and match with conclusions drawn from dashboard.

**RESULTS**

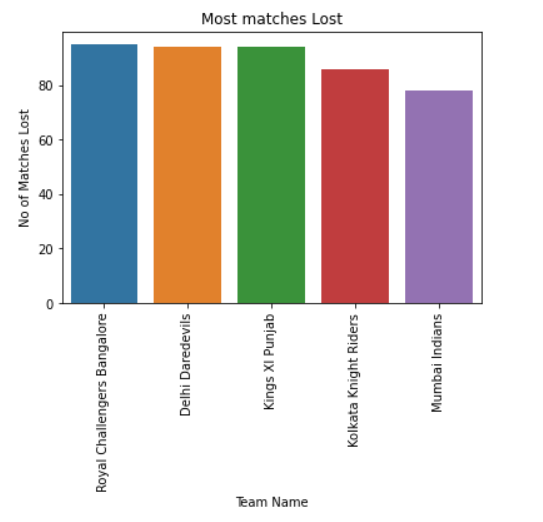
**1. To find the team that won the greatest number of matches in the entire IPL.**



**Answer**: Mumbai Indians

It can be observed that Mumbai Indians has won the highest number of matches i.e., **109**.

**2. To find the team that lost the greatest number of matches in the entire IPL.**

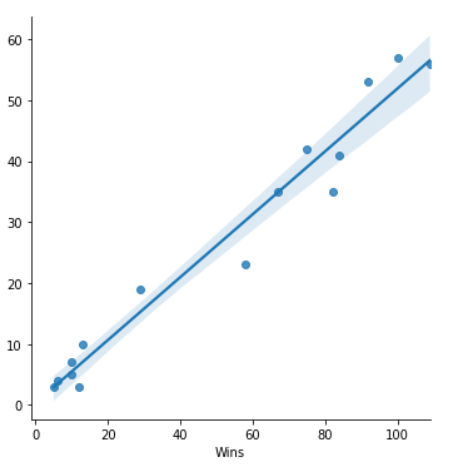
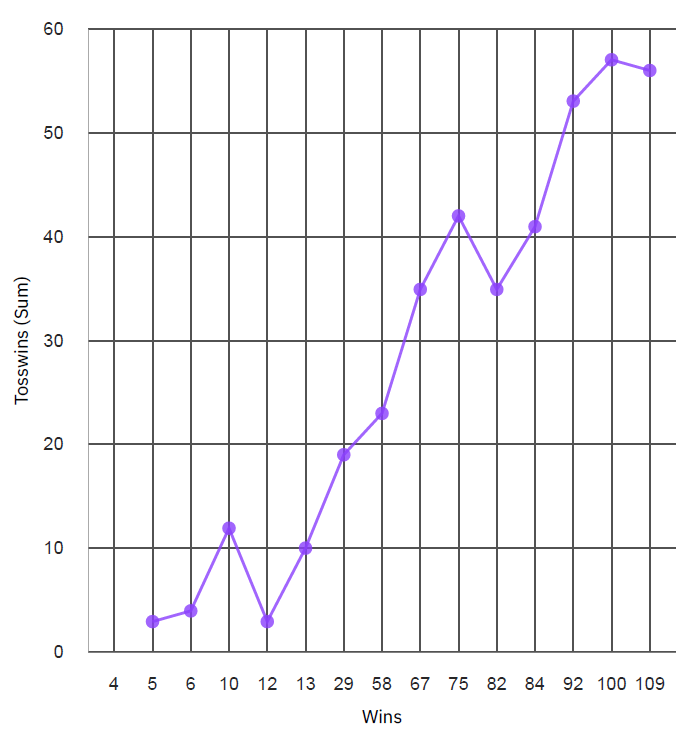


**Answer**: Royal Challengers Bangalore

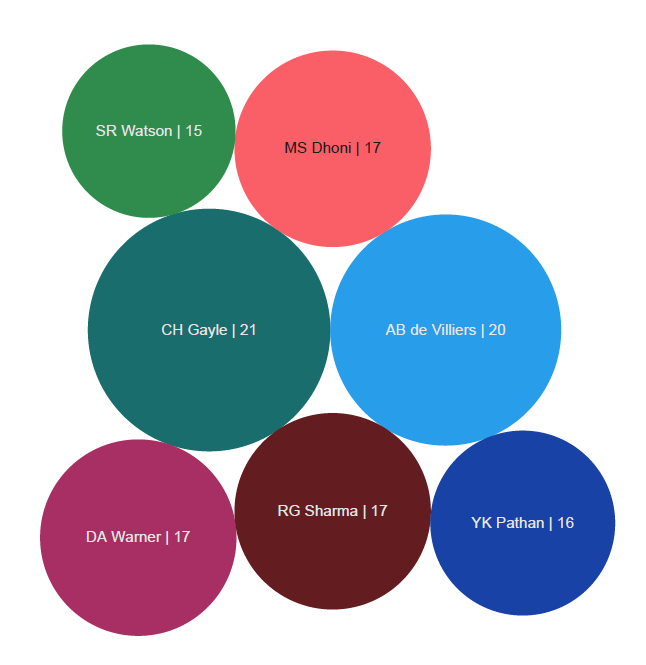
It can be observed that RCB has the highest Loses at **95** matches.

**3. Does winning a toss increase the chances of victory.**

**Answer**: The trend observed shows a linear relationship, but there are few outliers too. But it can be concluded that it does increase the chances of victory.

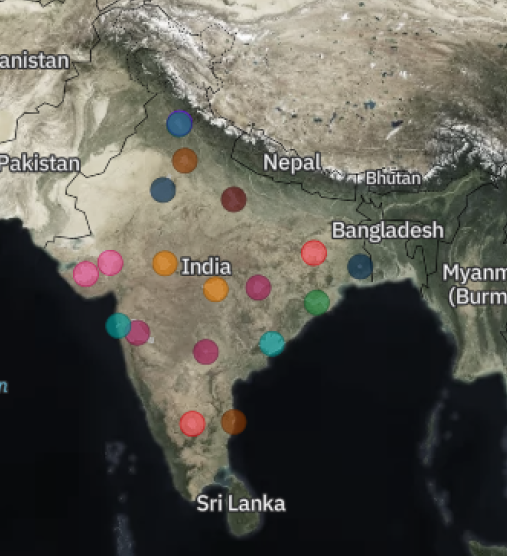
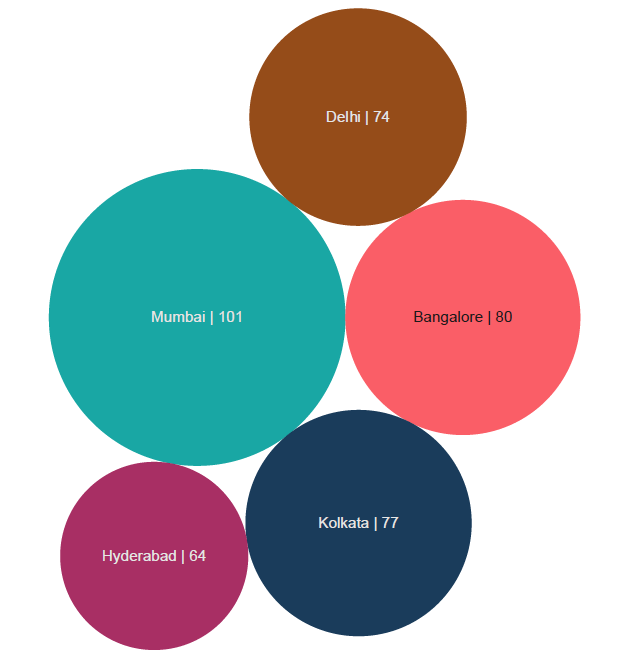
**4. To find the player with the most player of the match awards.**



**Answer**: CH Gayle

CH Gayle has most player of the match awards at total of **21**.

**5. To find the city that hosted the maximum number of IPL matches.**

**Answer**: Mumbai

Mumbai can be seen to host the greatest number of matches at the total count of it being **101**.

**6. To find the most winning team for each season.**

**Answer**:

**Season Teams No of Wins**

2008 Delhi Daredevils 10

2009 Delhi Daredevils 10

2010 Mumbai Indians 11

2011 Chennai Super Kings 11

2012 Kolkata Knight Riders 12

**Season Teams No of Wins**

2013 Mumbai Indians 13

2014 Kings XI Punjab 12

2015 Chennai Super Kings 10

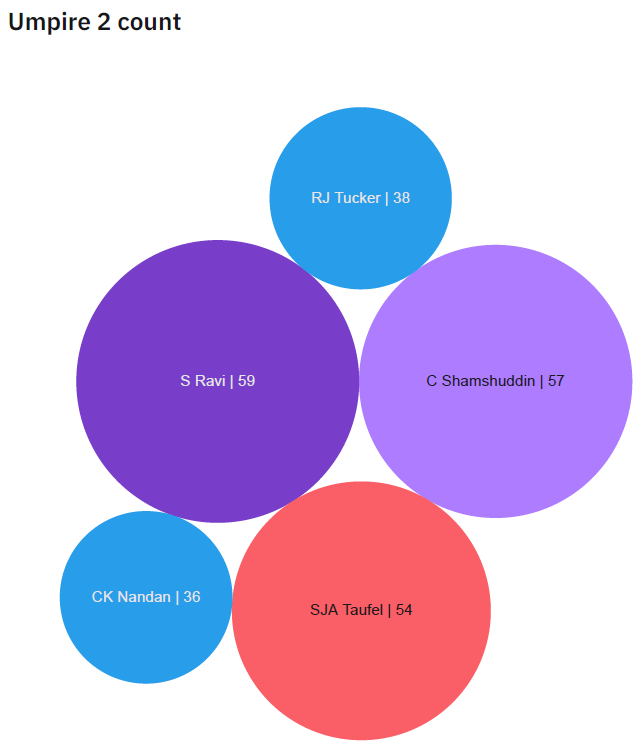
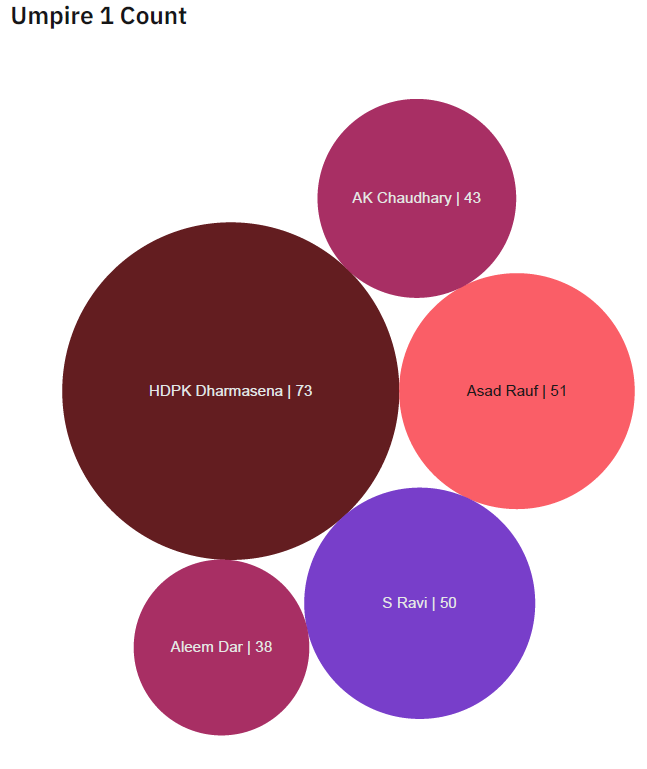
2016 Sunrisers Hyderabad 11

2017 Mumbai Indians 12

2018 Chennai Super Kings 11

2019 Mumbai Indians 11

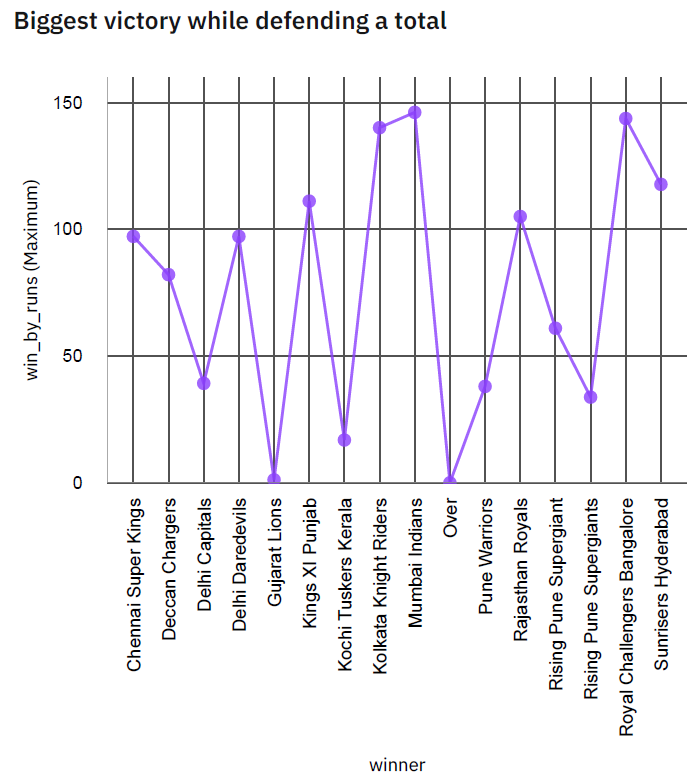
**7. To find the on-field umpire with the maximum number of IPL matches.**



**Answer:** S Ravi

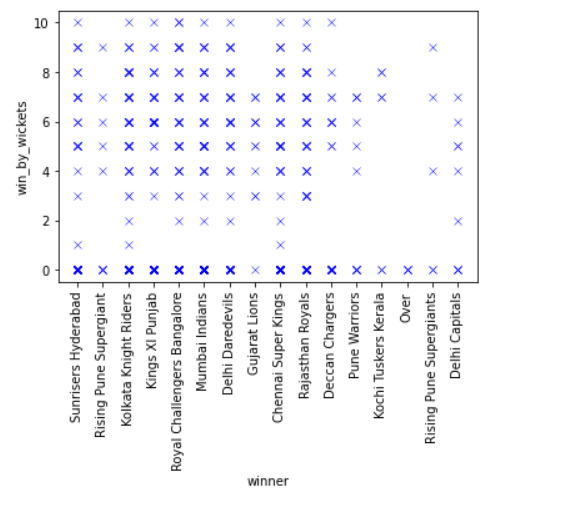
S Ravi name appears in both the column plots of Umpire1 and Umpire 2. Thus, it can be concluded the on-field umpire with most number of matches with a count of **109**.

**8. To find the biggest victories in IPL while defending a total and while chasing a total.**



**Answe**r:

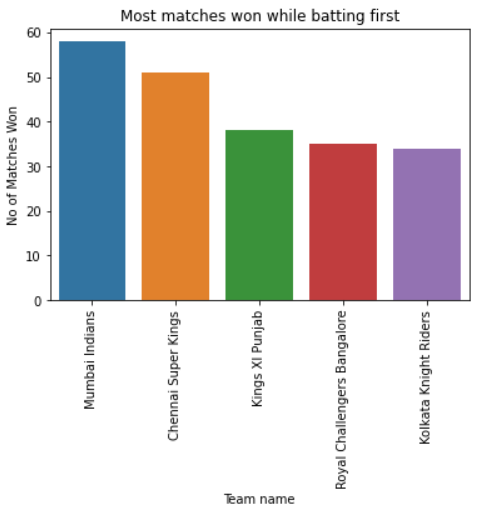
**While defending a total:** Mumbai Indians by **146** runs.



**Answer:**

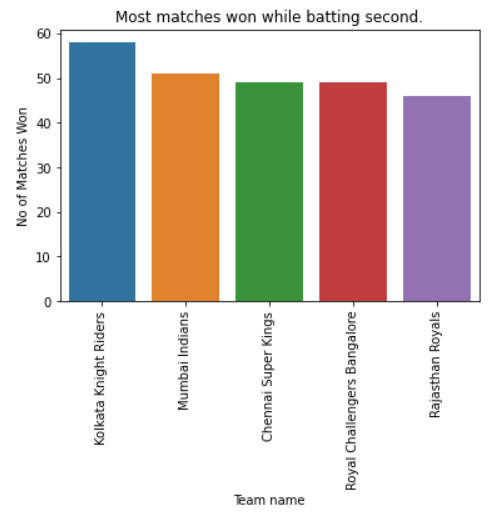
**While chasing a total:** Every team has atleast one biggest victory match. Conclusion arrived from the data. Biggest victory while chasing a total is when win by wickets count is zero.

**9. Which team won the most matches while batting first.**



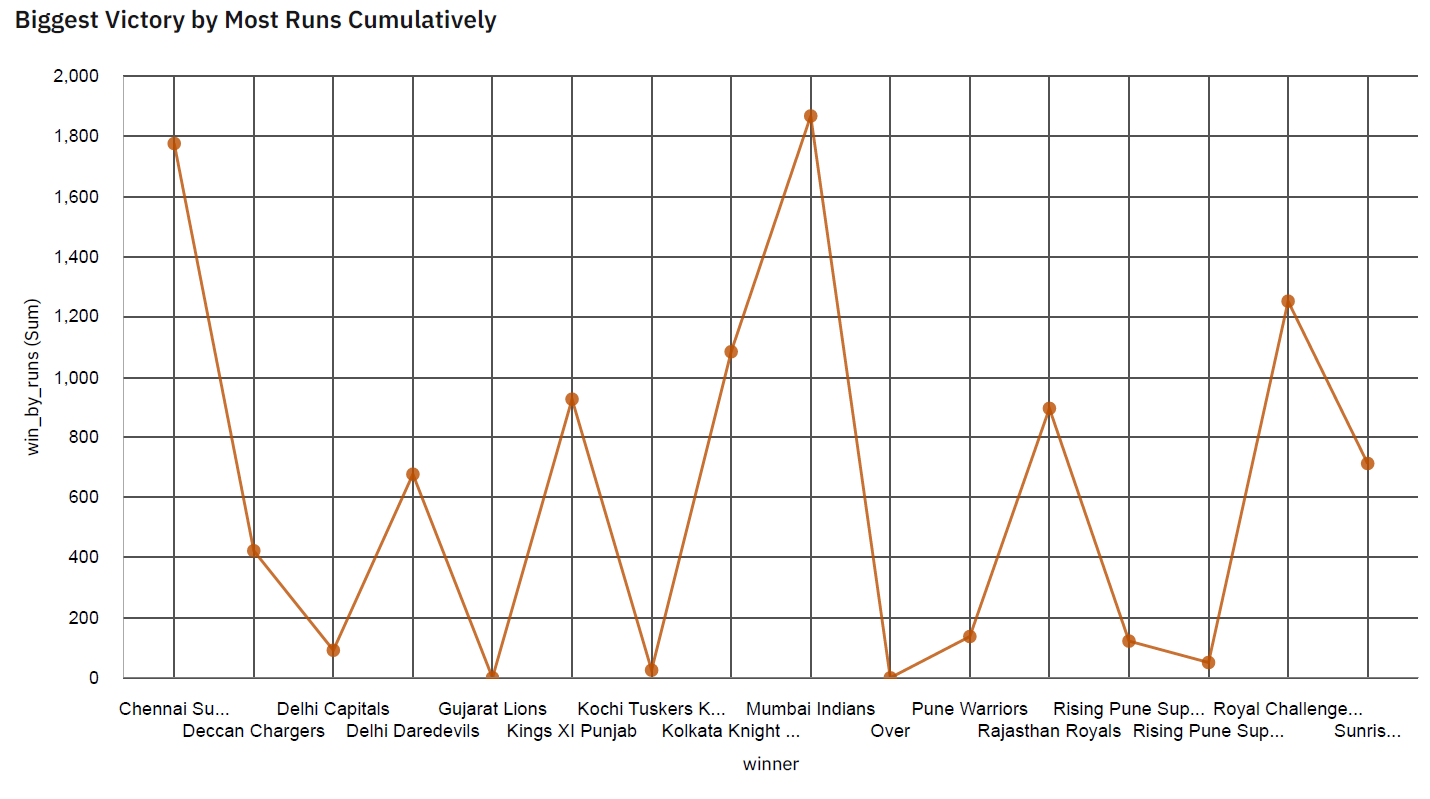
**Answer**: Mumbai Indians has the most wins while batting first with a total count of **58**.

**10. Which team won the most matches while batting second**



**Answer**: Kolkata Knight Riders has the most wins while batting second with a total count of **58**.

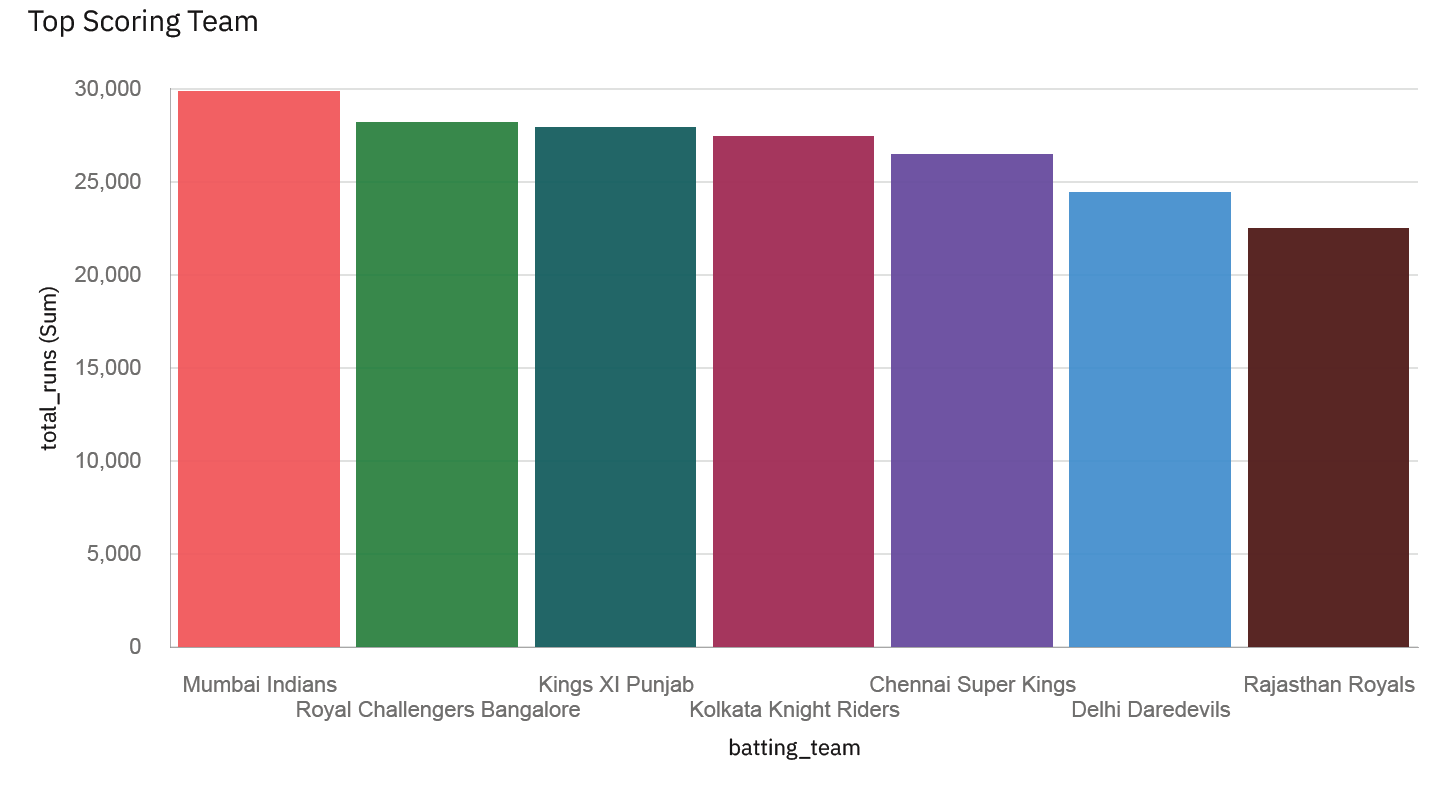
**11. List of teams which have won matches by most runs cumulatively.**



**Answer**: Mumbai Indians with **1866** Runs has the record of winning matches by most runs cumulatively.

**TEAM ANALYSIS:**

Here we find out the top scoring team having.

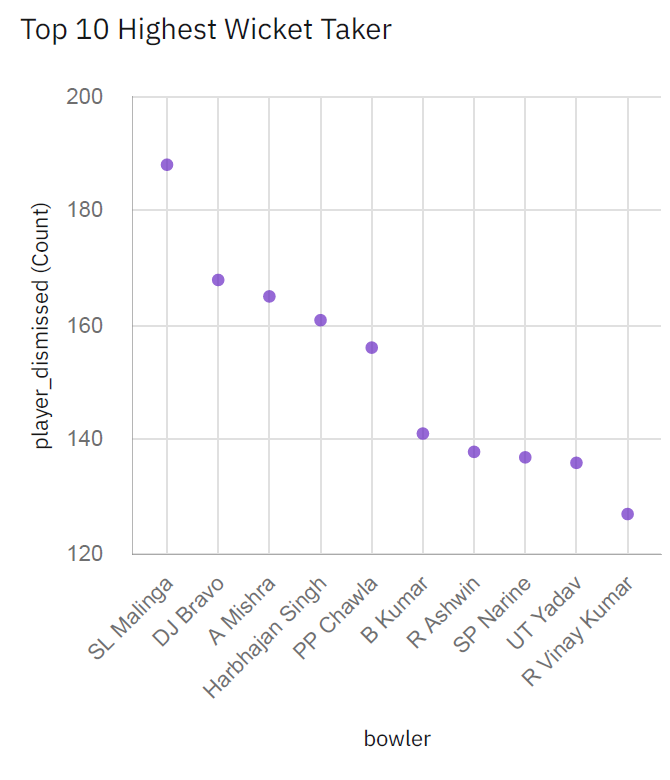


**Mumbai Indians** is the highest scoring team.

**PLAYER ANALYSIS:**



From the data it can be observed that the highest scoring batsman is **Virat Kohli**.



From the data it can be observed that the highest wicket taker is **SL Malinga**.

**Important Links**

**IBM Dashboard –**

Dashboard 1:

<https://ap1.ca.analytics.ibm.com/bi/?perspective=dashboard&pathRef=.my_folders%2FIPL%2BSuper%2BPredictor&action=view&mode=dashboard&subView=model0000017b73df0d29_00000000>

Dashboard 2:

<https://ap1.ca.analytics.ibm.com/bi/?perspective=dashboard&pathRef=.my_folders%2FIPL%2BTeam%2Band%2BPlayer%2BAnalysis&action=view&mode=dashboard&subView=model0000017b7759c99e_00000000>

**Video Submission –**

<https://drive.google.com/file/d/1TbMuhAQhRZrWMiVCEM4C7-XvK0nIIaQi/view?usp=sharing>

THANK

YOU