DSCI 590 Fall 2023

Indian Premier League Data Visualization

Harsha Valiveti, Sahith Thupakula

Abstract

Cricket is a prevalent sport that is played all over the world, though in some countries, such as India and Pakistan, it is more popular. With the number of matches increasing by the day, it has become challenging to manage and extract useful information from all of the available data. Many teams use this data to improve their performance by interpreting and analyzing statistics in order to gain a competitive advantage over their rivals. All of this data can also be used to enhance understanding and management of the fan experience. We collected IPL (Indian Premier League) data for this study and used various Python libraries and predictive models to visualize the performance of the players and teams. In addition, the data is used to create visually appealing and informative visualizations of the overall outcomes of a cricket league.

Introduction

Cricket matches are played in a variety of formats, including One Day Internationals, Twenty20 matches, and Test matches. The Indian Premier League (IPL) is a Twenty-20 cricket tournament league that was founded to promote cricket in India and thus develop young and talented players. An annual event in which teams from various Indian cities compete against one another is known as a league. The Board of Control for Cricket in India (BCCI) founded it, and it has since grown into a massive, lucrative cricket venture. The IPL teams are chosen through an auction. Player auctions are not a new phenomenon in the sports world. However, in the Indian Premier League, for the first time, a team was selected from a pool of available players via player auctioning (IPL). Due to the involvement of money, team spirit, city loyalty, and a massive fan following, the outcome of matches are extremely important for all stakeholders. This is determined by the game's complex rules, the team's luck (Toss), the ability of players, and their performances on any given day. Other natural parameters, such as player history, are crucial in predicting the outcome of a cricket match. Predicting the outcome of games between different teams can be beneficial to the team. The primary goal of this project is to visualize this cricket data. It's fascinating to learn how popular a sport is in a particular area, and we can also choose which team to support based on their success. By looking at these, investors can learn which teams to invest more heavily in from a financial standpoint, and team management can learn which teams to invest more heavily in from a tactical standpoint. Top international sports leagues, such as the IPL and the T20 World Cup, have expanded their operations not only in their home countries but also globally. Many visualizations have been created by media and cricket analysts. Every match is instantly visualized and broadcast on television in between games.

Process

The first data we had was a ball-by-ball record of all formats of cricket from various years. We first filtered out the IPL format data. The filtered IPL data includes information such as the year, the innings being played, the teams playing that particular inning, all the teams in that particular season, the score of each ball, the record of who the bowler and batsman are

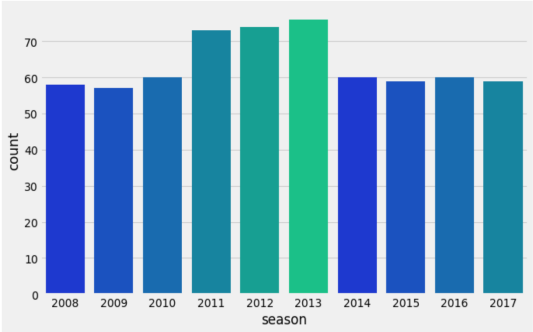
for that particular ball, the stadium where the match is being played, who won the match, who won the toss, who was the captain for the team at the time, and so on. The dataset had approximately 1983 columns at first, we preprocessed it to remove any duplicates. The NaN values were removed and replaced with 0. We converted the categorical variables to numeric variables and removed all the columns that had little correlation and provided little information. After all of this pre-processing, we had clean, structured data to work with.

For visualization, the python packages NumPy, pandas, Matplotlib, Seaborn, and Plotly were used. We also created a few visualizations using the basic tableau knowledge we gained during the process. There were a few failed visualizations along the way that didn't show many trends in the plot and had little to interpret from the plot. The concept of correlation was then implemented and applied here to understand which features have a better interaction that can be visualized. We encountered a few challenges during preprocessing. For example, we discovered that having 4s and 6s in separate columns did not result in the expected interaction with the rest of the data. So we combined everything into a single column called "scores." This column contains the score of that specific ball. Then, whenever we needed to find the total number of 4's or 6's in that innings, we made use of Python's count method. It wasn't just the scores; many other columns had to be dropped or merged to form more structured data.

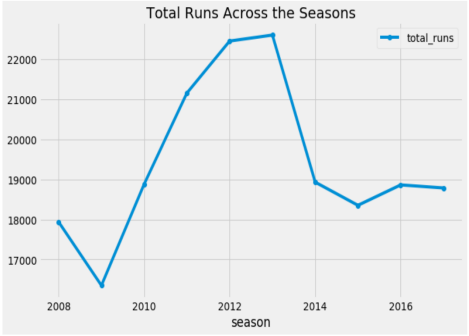
We ensured that the data set was structured and clean throughout and that plots were not manipulated due to false values. To improve our visualizations, we also used the best design and color principles. We presented the data with precise labeling, and the data-intensive visualization was adjusted to highlight trends.

Results and Insights:

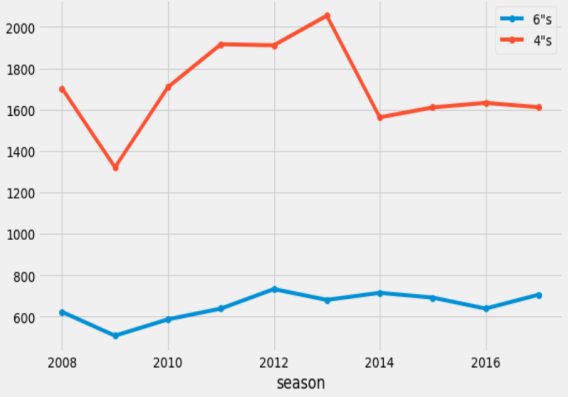
From the available visualizations, we have gathered some insights about IPL, of all the seasons, 2013 was the biggest in terms of matches held.



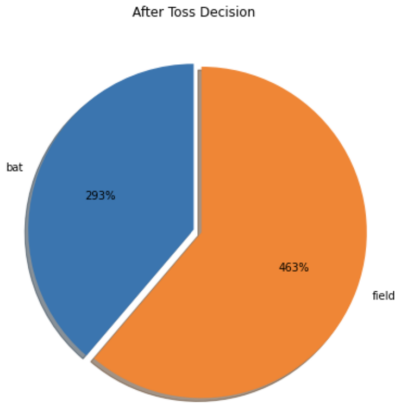
As 2013, even 2012 has the highest number of IPL matches organized, the total runs scored by all teams at that time was 23000 and 22500, which was the biggest.



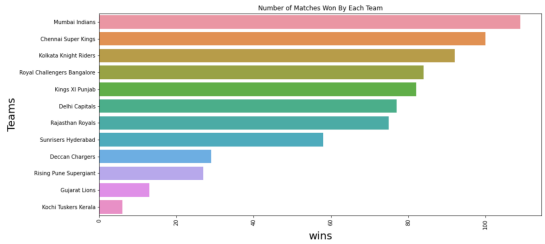
As the match happens, players score big with the boundaries and score high, of all matches that happened, 2012 has the most number of sixes closing to 800, and 2013 has the highest boundaries of 2000 fours which is tremendous.



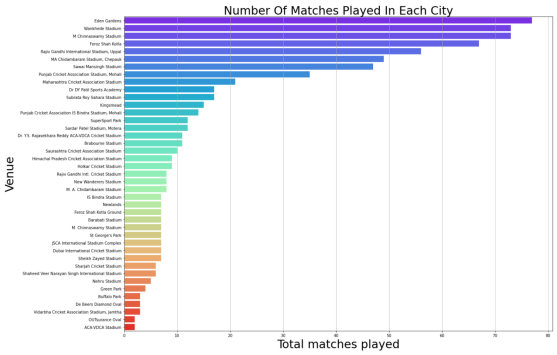
Looking at the data, we tried some interesting visualizations, The match starts with a Toss, Toss winner will decide what to choose Bat/field, the insight shows which decision impacted the win, if a team wins the Toss and choose to field, showed their likeness to win with a percentage of 46% and for the field resulted in the win by 29.3%.



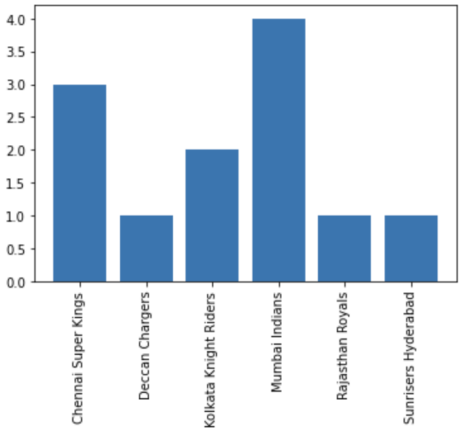
The next decision is which team won the most number of times, it shows that Mumbai Indians won 109 times, followed by Chennai Super kings with 100 wins.

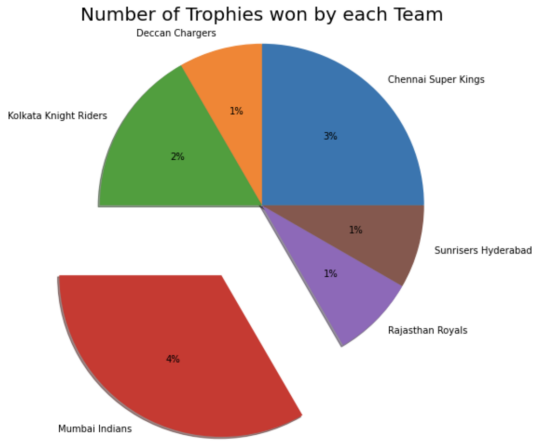


The match is hosted in different cities, the most number of times hosted in Eden Gardens, which hosted for 77times.

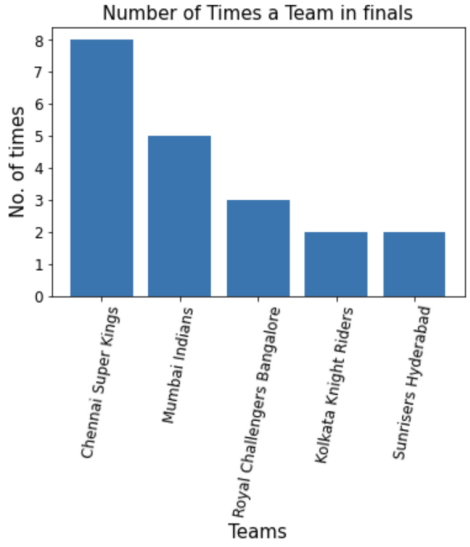


Who won the most number of final titles in a season? The Team Mumbai Indians won the 4 trophies, next Chennai Super Kings with 3 trophies.

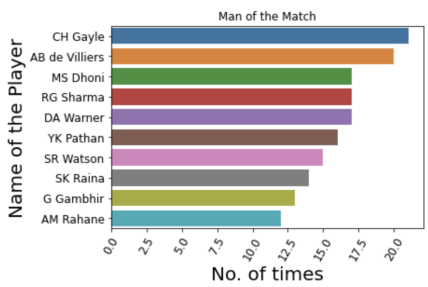


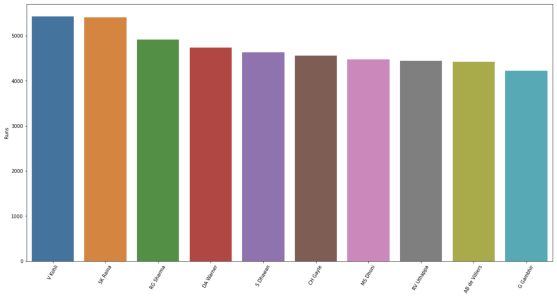


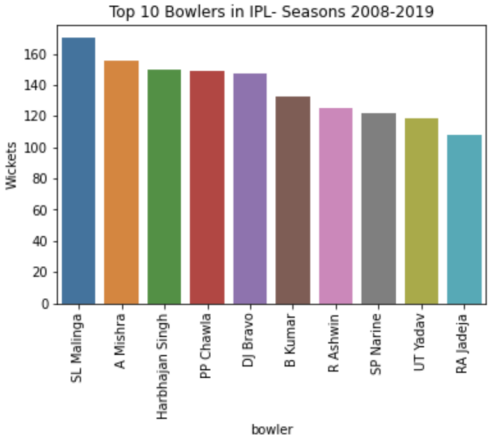
To understand the competition, in entering the finals, Chennai Super Kings, which played 8 times, noticeably won the final cup by 3, which is impressive.



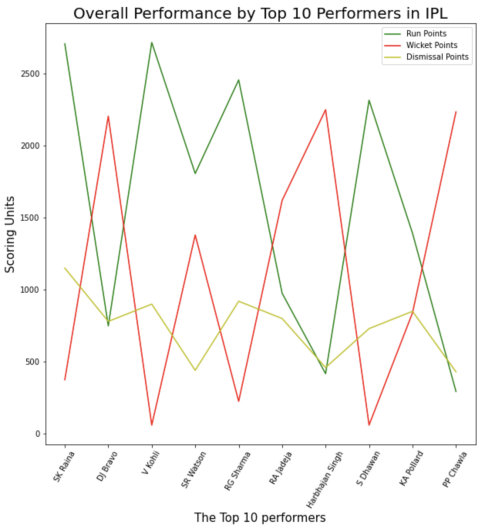
In the Player analysis, in each season at the end of a match they announce the man of the match, the data showed Chris Gayle won the Man of the Matches with 21 titles.



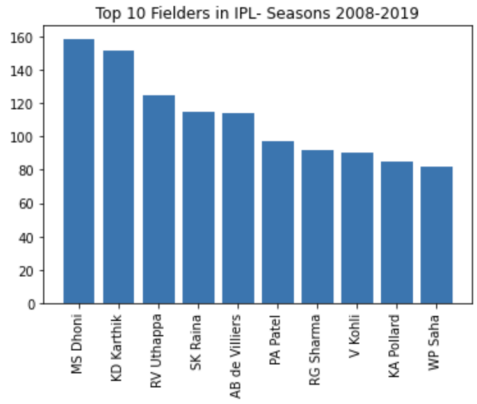




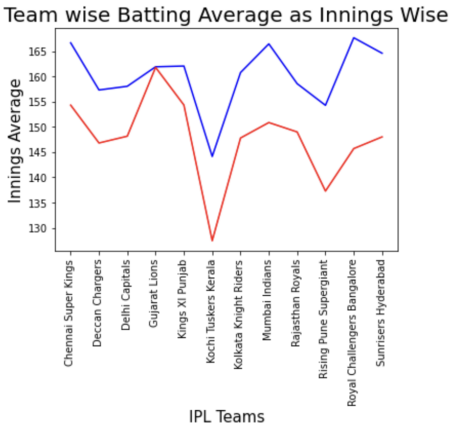
The best all-around performance by a player with good batting, bowling, and fielding will assess the player. Suresh Raina has scored all of them in the IPL.



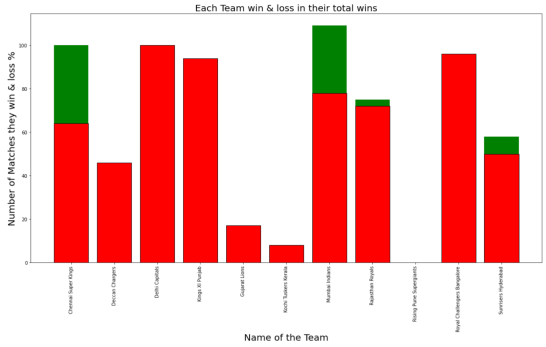
The most number of catches taken by a player in the IPL is MS Dhoni with 159 catches.



For any team, their performance depends on the score they put in, in the first innings and the second innings. Gujrat Lions played consistently and maintained their scores at an average of 160 runs in both innings.



What are the team wins against their losses? Most times when a team plays, they saw a defeat when compared to their wins, Chennai Super Kings is the one which has more difference with win and loss, which is evident that this team has been performing consistently and became everyone's favorite team.



Discussion:

From the visualizations, after the Toss, how many teams won after taking a decision (bat/bowl) the most number of wins recorded by taking bowling i.e. fielding is 46% and the number of wins recorded by taking batting is 29%. Mumbai Indians are the team that has won the IPL trophy the most. Mumbai Indians defeated Delhi Daredevils with the highest run difference. Kolkata Knight Riders defeated Gujrat Lions with the highest wins by wickets. The decision for batting or fielding varies largely across the seasons. In some seasons, the probability that tosses winners opt for batting is high, while it is not the case in other seasons. In 2016 though, the majority of toss winners opted for batting. Mumbai Indians seem to be very lucky to have the highest win in tosses followed by Kolkata Knight Riders. Pune Supergiants have the lowest wins as they have played the lowest matches also. This does not show the higher chances of winning the toss as the number of matches played by each team is uneven.

Between 2008 and 2009, fewer runs were completed overall. But following that, runs significantly increased every season up until 2013, after which there was a decline in total runs. But not every season has the same number of games. The average runs per game for each season should be examined. The final five overs of the game account for the majority of runs scored. The number of runs scored by MI and RCB has been rising

throughout the game. In 13 of the 18 games that have been played between the two teams, MI has won. Only in 2014 did KKR triumph in both games. Thus, we know who to bet on in a MI vs. KKR encounter. Any two teams can be compared in a similar way; all we need to do is swap the team names. One thing to note is that MI and KKR have never faced off against one another in a qualifying round or championship game because they each only participate in two matches each year, which are group-stage games.

Another interesting question is whether any team scoring high runs above 200 runs and batting first will lead to a win, yes! A team batting first and scoring 200+ runs resulted in victory. All seasons combined, Virat Kohli has scored the most runs. Only a few runs separate Raina from second place. David Warner's performance seemed to be getting better every year. Kohli's runs have significantly decreased throughout the past season. It appears that the winning team should choose to bat after winning the toss because, out of 9 finals, batting has won matches 4/5 of the time. Chris Gayle is far ahead and has the maximum sixes. Chris Gayle tops the list with a score of 175 and the most centuries counting to 5. An interesting factor is, only a single player who played constantly for a team in all seasons starting from 2008 to 2022 is Virat Kohli.

Conclusion:

With the help of these visualizations, one can glean important insights from historical data and use those conclusions to make informed decisions. They can also help with efficiency by allowing people to look back on their actions and results in the past. Through this project, we have learned about plethora of python libraries that can be used to make beautiful and informative visualizations. We processed and structured data making using of best techniques. We paid special attention while making color and design choices for the visualizations so that the plots are interesting and easy to understand. In the future, we can adapt the same for different formats of cricket and also different sports like soccer, badminton, the Olympics, or any sports data. We want to enhance on these visualization project, craft powerful visulization that can be used by executives and common sports enthusiasts. We also plan to use these visualization to build interactive live dashboards in the future for easy access and interactivity.

References

1. https://www.kaggle.com/datasets/nowke9/ipldata
2.https://towardsdatascience.com/exploring-indian-premier-league-with-interactive-visualizations-7a6ae053449
3.https://www.kaggle.com/code/paramarthasengupta/eda-indian-premier-league-2008-2019/notebook