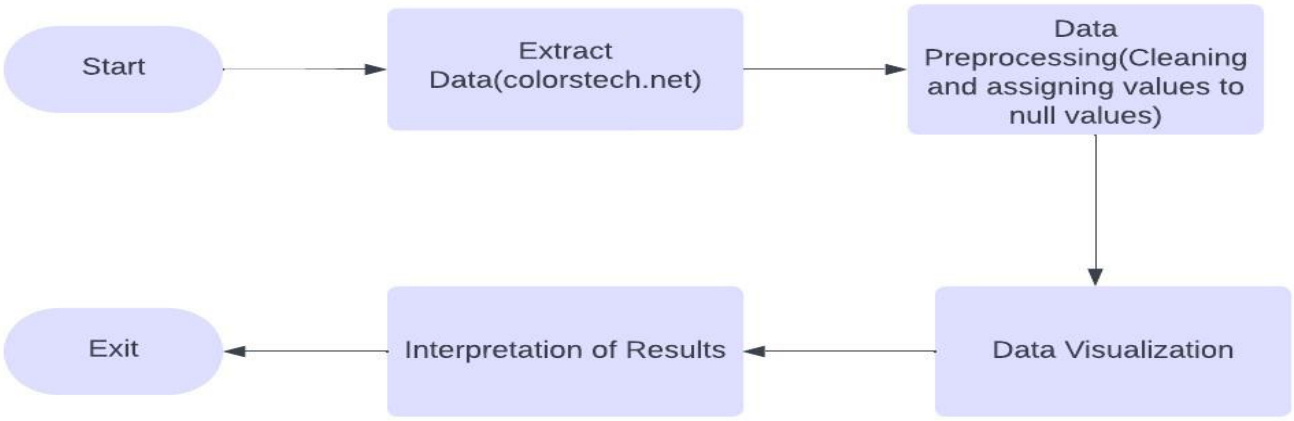
Exploratory Data Analysis of IPL season-wise Data

NOTE: THE DATA IS ONLY TILL 2023 WHICH IS THE CURRENT LATEST SEASON OF IPL AT THE TIME OF ANALYSIS

Exploratory Data Analysis (EDA) is a foundational step in the process of data driven decision making or predictive modelling, used across various studies to uncover patterns, trends, and relationships within datasets before conducting more in-depth statistical analyses. The essence of EDA lies in its ability to inform further research directions through the identification of data distribution, outliers, and associations among variables. This process often utilizes visualizations, summary statistics, and outlier detection to gain insights that guide further analysis The process diagram for sports analytics is illustrated below:



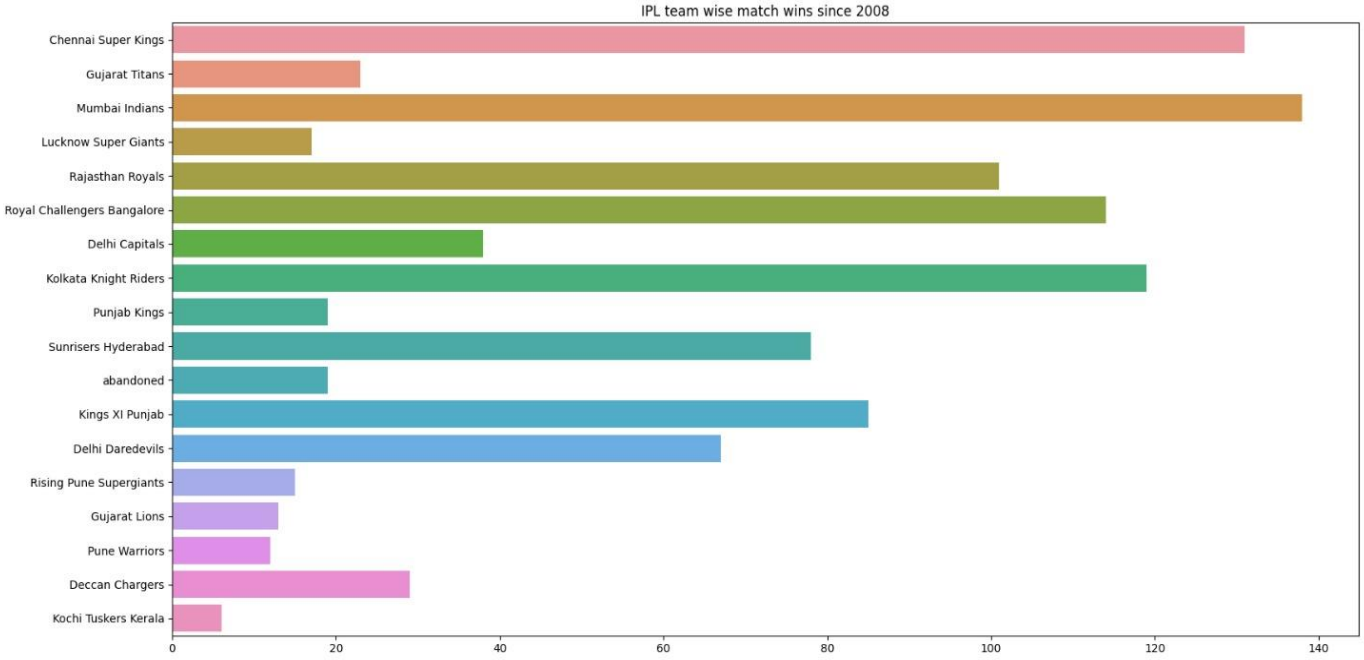
Data Pre-processing

Data pre-processing is a foremost step in the data analysis process, especially in sports analytics, where the quality and format of data can significantly affect the outcomes of the analysis. The main goal of data pre-processing is to prepare the raw data for analysis by making it clean and structured. Data Cleaning is a vital step in the data pre-processing phase and focuses on identifying missing values, detecting and removing outliers, filtering or removing unwanted data, correcting data inconsistencies, and handling erroneous data.

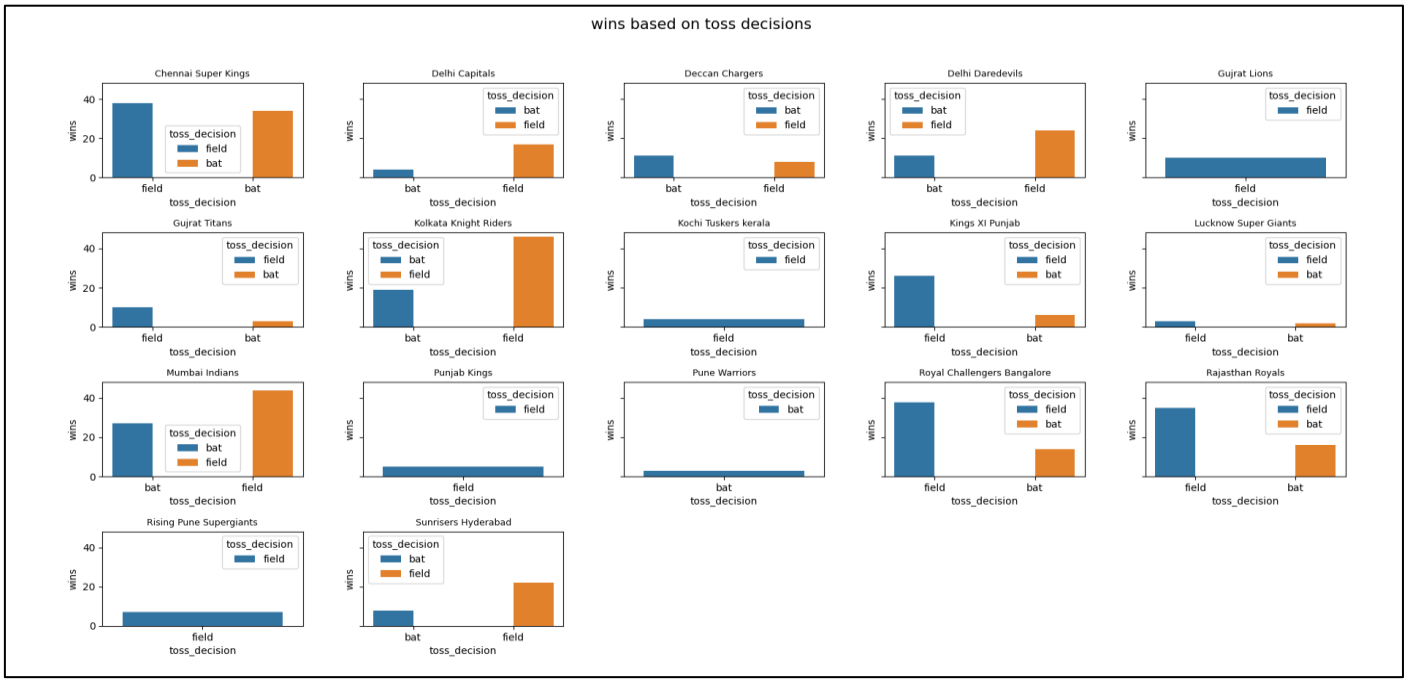
- Dropping unwanted columns: In the first step, we drop the unwanted columns from the dataset which are not required for the analysis. For this analysis, we drop columns namely id, city, date, toss_winner, result, dl_applied, win_by_runs, win_by_wickets, and umpire.
- Identifying discrepancy in Data: In this step after discarding all the unnecessary features, we take an overview of the data and it was found that the dataset was missing records for 7 matches. To identify and point out this discrepancy the count for all records in the dataset was taken and matched with the official data on the Indian Premier League.
- Adding missing data records in the dataset: in this step we append the data for the 7 missing matches identified in the previous step and after matching the data to the data provided by the IPL, manually, it was found that all the missing matches were matches that were abandoned without a single ball being bowled.
- Assigning values to null values: Assigning values to null values in a dataset is a crucial step for aiming to maintain data integrity and improve the quality of analysis. This process, known as imputation, involves replacing missing values with replaced values based on various methods. In our dataset, we are applying constant value imputation to replace the null values in the" winner" column with "abandoned".
- Replacing duplicate names of stadiums having minor differences, with a single one in order to get correct analysis of topics pertaining to the venue, as minor changes in name can affect the plots that are drawn.

Data Visualization

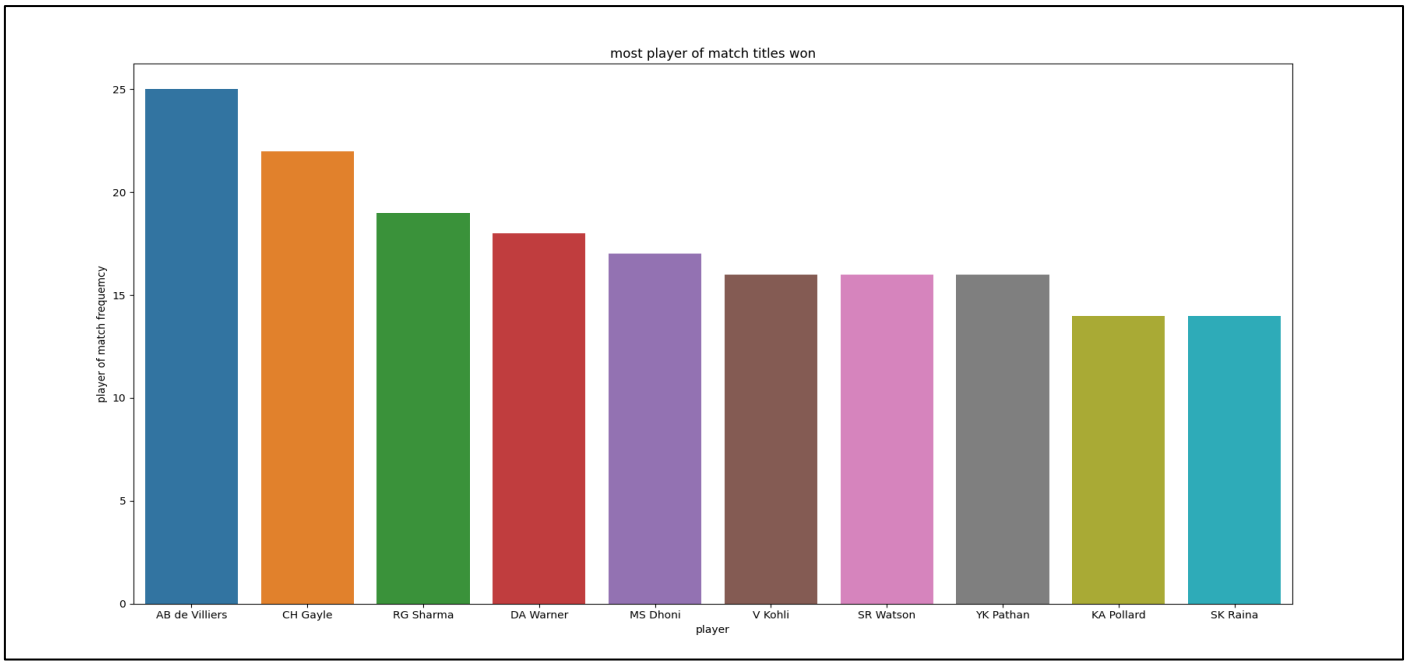
The overall sports analytics for IPL has been visualized with Python for 15 years. In this manuscript, we are using two main libraries of Python Matplotlib and Seaborn. Matplotlib provides the foundation for constructing detailed plots, Seaborn offers an enhanced interface that simplifies the process of creating sophisticated statistical visualizations.



Analysis: in the above graph we can see a visual representation of the total matches won by each team that has partaken in the Indian Premier League since its inception in 2008 till the latest 2023 season available in our dataset. The plot depicts the no. of wins on the Y-axis and the teams on the x-axis with a different hue for each team. The graph conveys Mumbai Indians being as one of the more successful team with a win count of about 138 wins throughout all the seasons that it has played in. On the other hand it can be seen that Kochi Tuskers Kerala have the fewest wins as the team was terminated after playing for just a single season in 2011.



In the graph seen above the aim is to visually depict and by extension analyse the dependence of toss decision on the number of matches won by each team. the disparity that can be seen in the graph numbers is a direct consequence of the changing dynamic of the Indian Premiere League (IPL) where some teams have been terminated and some have been renamed to their current names.



The above bar graph aims to depict the players that have the most **'player of match'** titles in the league, regardless of the team that the player plays from or season. the above bar graph has entries for the top ten players who had the highest frequency for the field **player_of_match**.

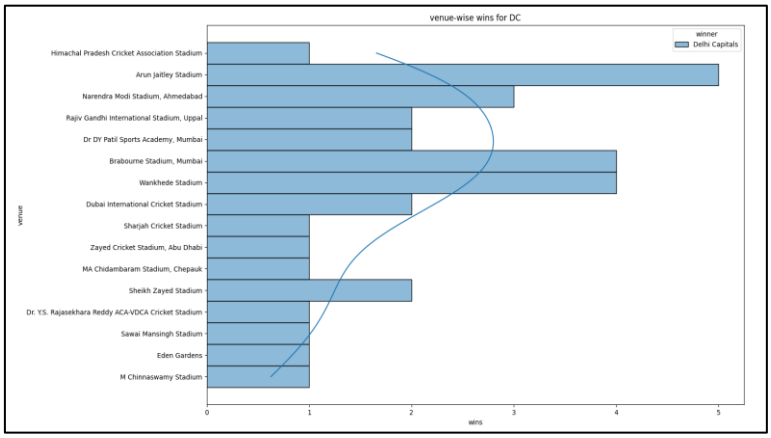


Figure 6(a)

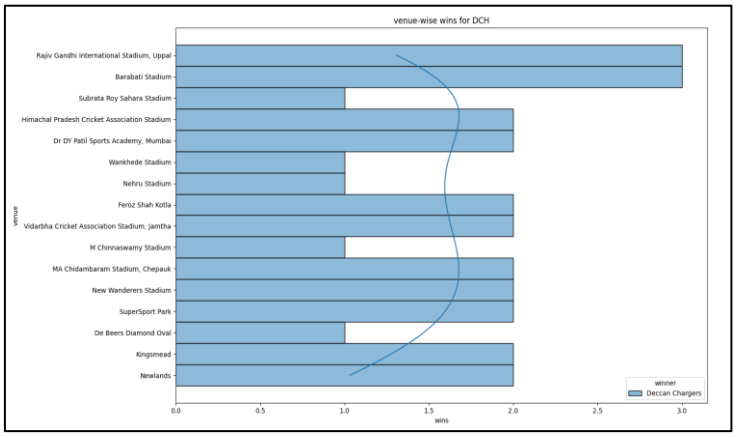


Figure 6(b)

Figure6(a) shows the wins for team Delhi Capitals on every venue on which they have ever played a match, similarly Figure6(b) provides for a similar plot of wins for the team Deccan Chargers on all venues that the team has ever played on since the start of the league.

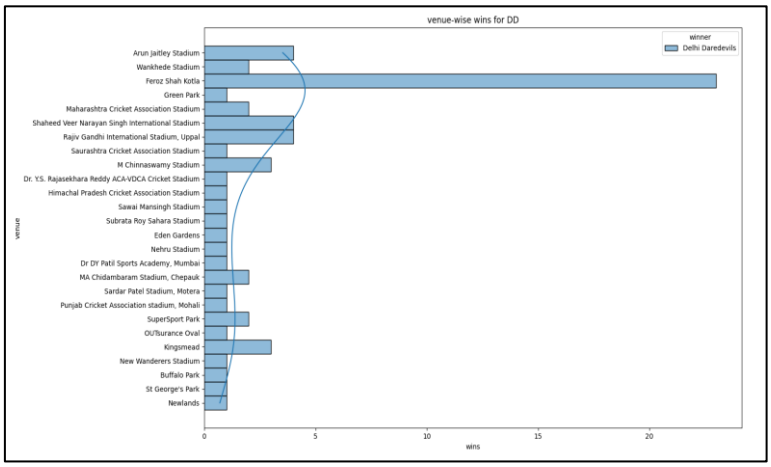


Figure 6(c)

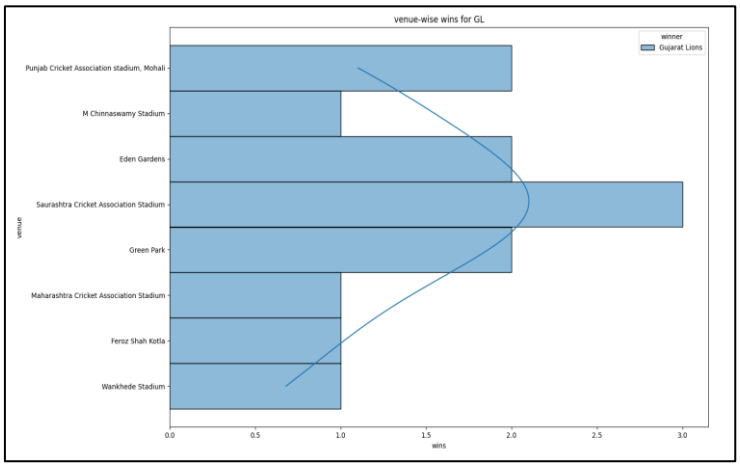


Figure 6(d)

Figure6(c) shows the wins for team Delhi Daredevils on every venue on which they have ever played a match, similarly Figure6 (d) provides for a similar plot of wins for the team Gujarat Lions on all venues that the team has ever played on since the start of the league.

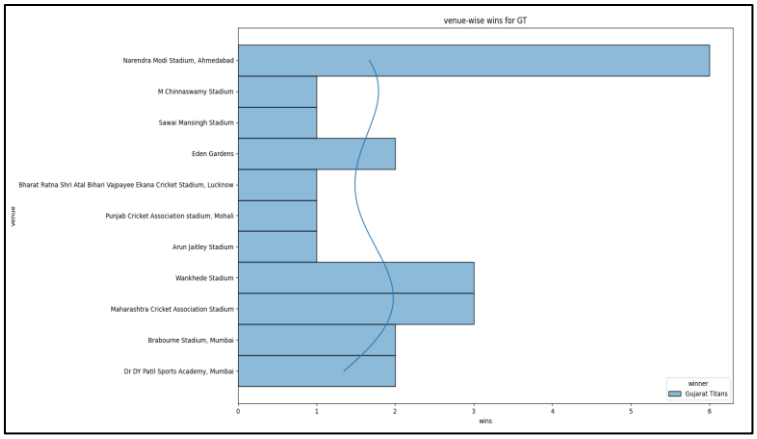


Figure 6(e)

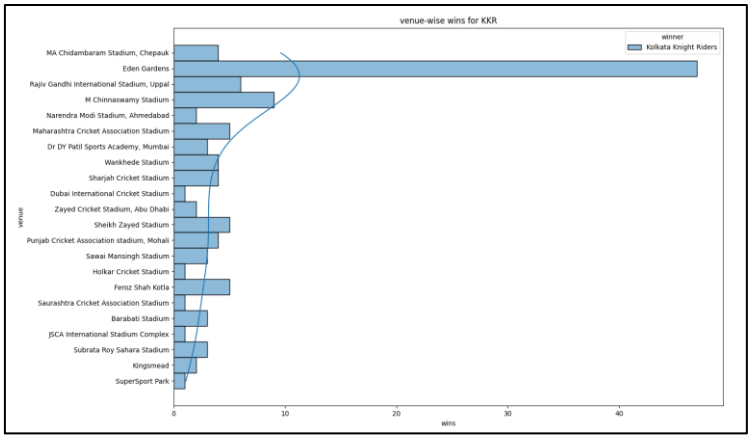
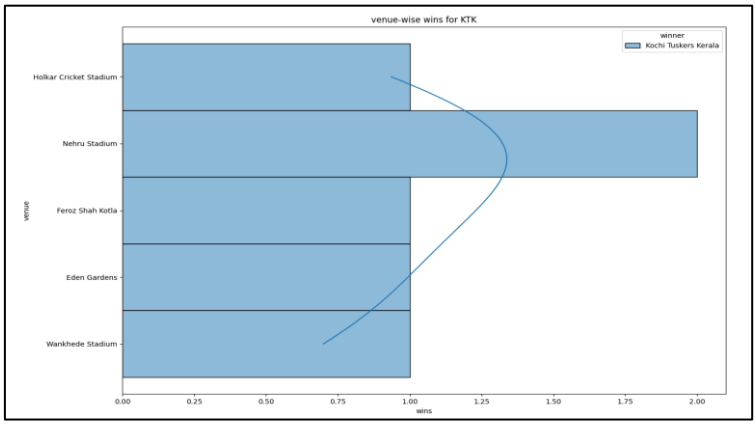


Figure 6(f)

Figure 6(e) shows the wins for team Gujarat Titans on every venue on which they have ever played a match, similarly figure 6(f) provides for a similar plot of wins for the team Kolkata Knight Riders on all venues that the team has ever played on since the start of the league.



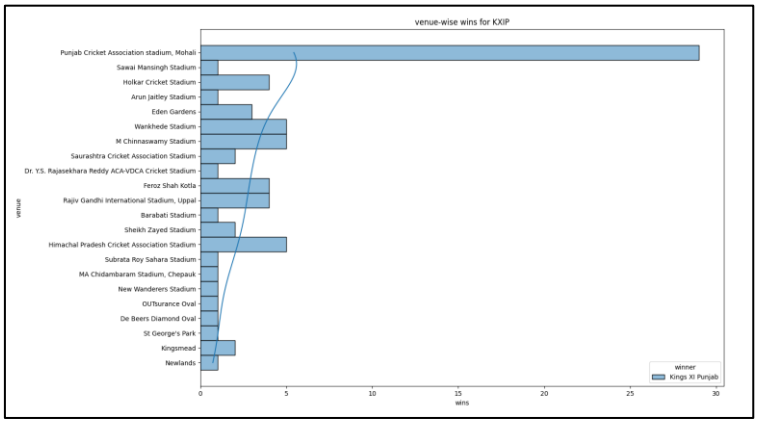
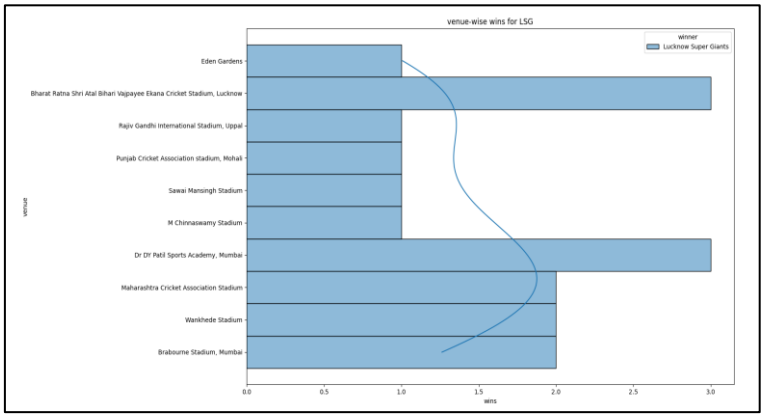


Figure 6(g) **Figure 6(h)**

Figure 6(g) shows the wins for team Kochi Tuskers Kerala on every venue on which they have ever played a match, similarly figure 6(h) provides for a similar plot of wins for the team Kings XI Punjab on all venues that the team has ever played on since the start of the league.



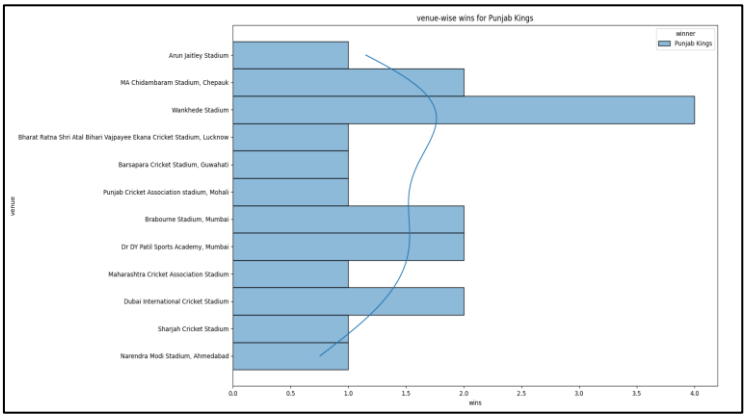
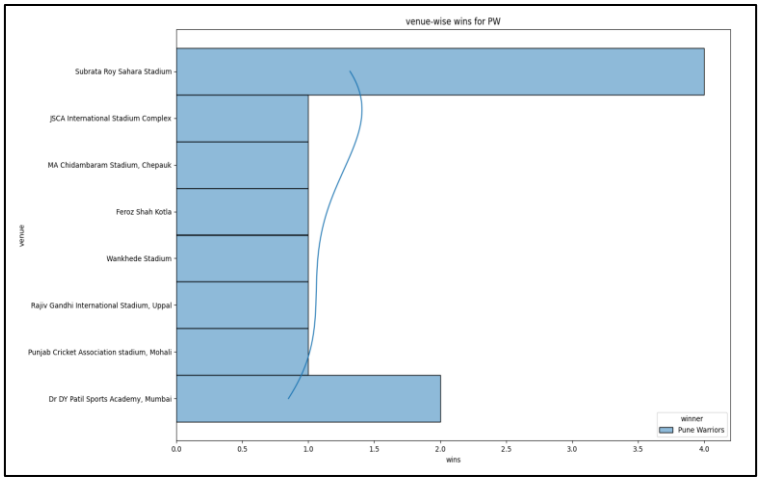


Figure 6(i) **Figure 6(j)**

Figure 6(i) shows the wins for team Lucknow Super Giants on every venue on which they have ever played a match, similarly figure 6(j) provides for a similar plot of wins for the team Punjab Kings on all venues that the team has ever played on since the start of the league.



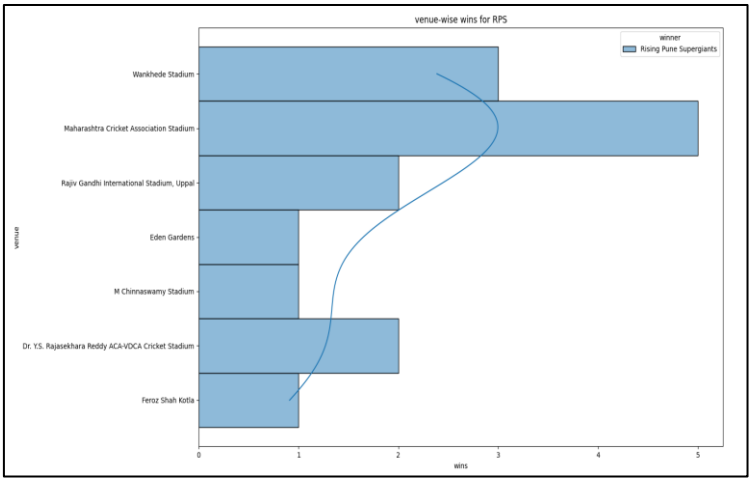


Figure 6(k) **Figure 6(l)**

Figure 6(k) shows the wins for team Pune Warriors on every venue on which they have ever played a match, similarly figure 6(l) provides for a similar plot of wins for the team Rising Pune Supergiants on all venues that the team has ever played on since the start of the league.

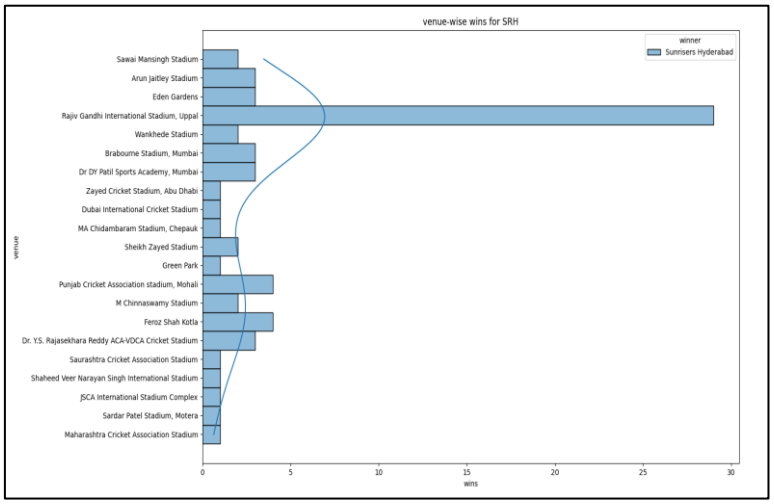


Figure 6(m)

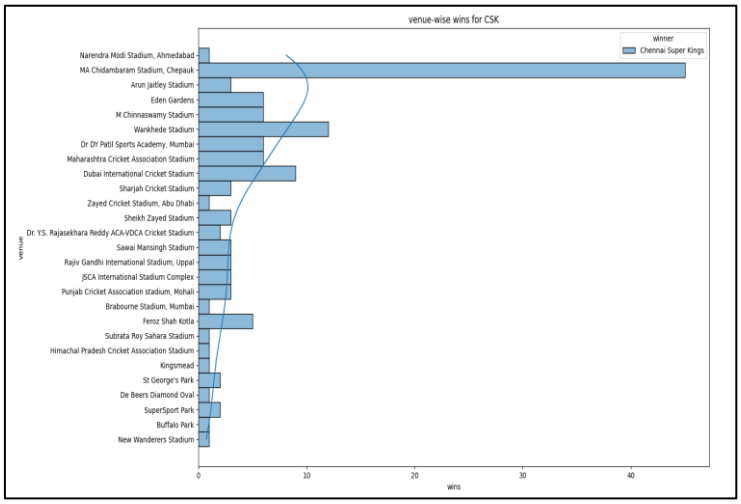


Figure 6(n)

Figure 6(m) shows the wins for team Sunrisers Hyderabad on every venue on which they have ever played a match, similarly figure 6(n) provides for a similar plot of wins for the team Chennai Super Kings on all venues that the team has ever played on since the start of the league.

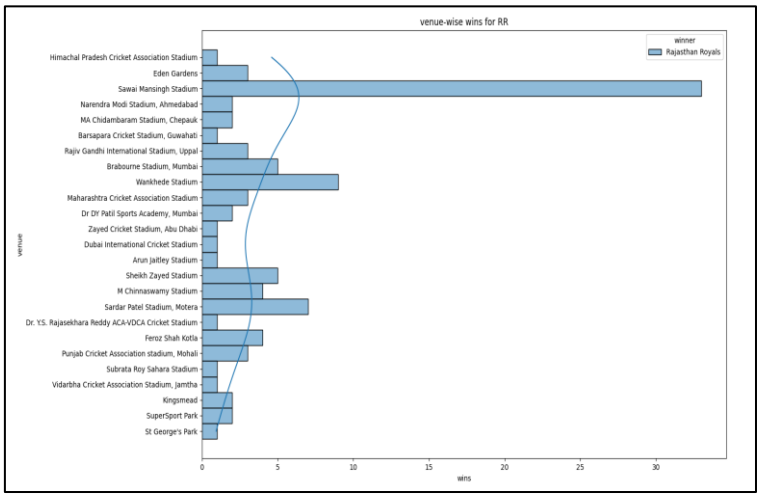


Figure 6(o)

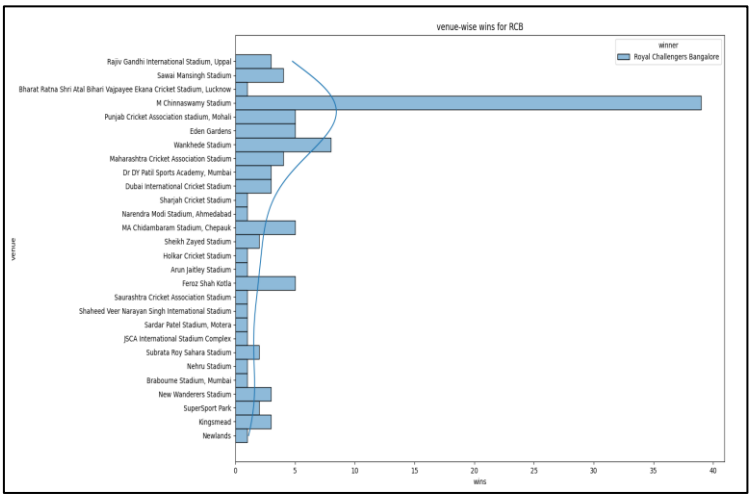


Figure 6(p)

Figure 6(o) shows the wins for team Rajasthan Royals on every venue on which they have ever played a match, similarly figure 6(p) provides for a similar plot of wins for the team Royal Challengers Bangalore on all venues that the team has ever played on since the start of the league.

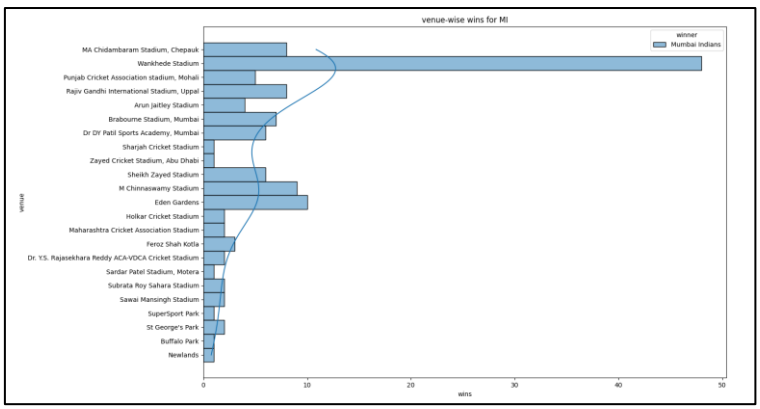
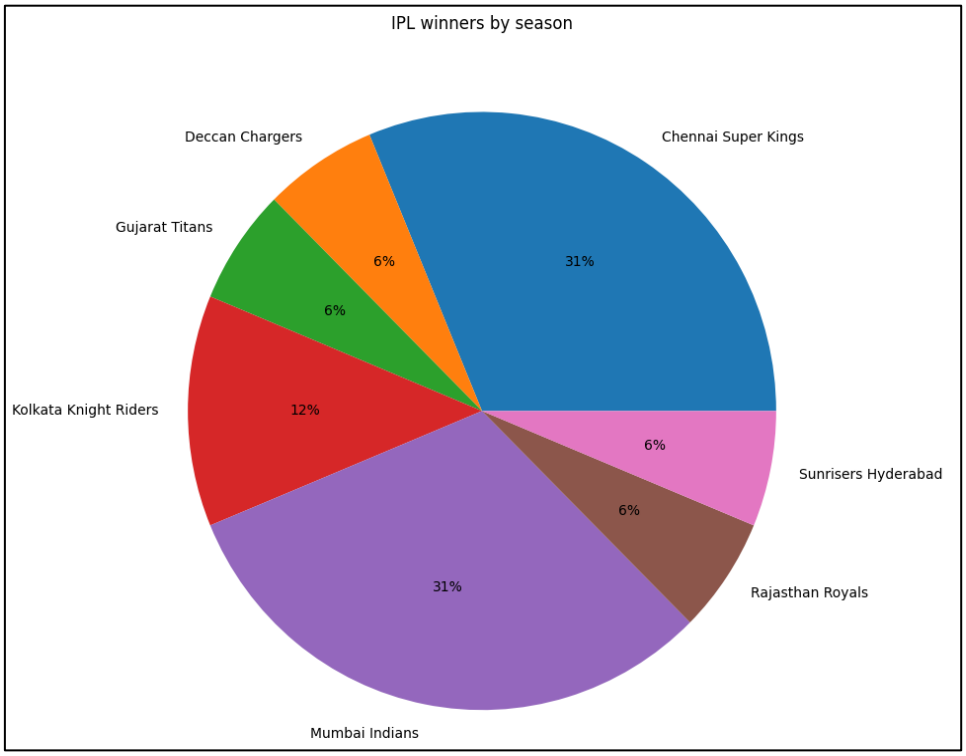


Figure 6(q)

Figure 6(q) shows the wins for team Mumbai Indians on every venue on which they have ever played a match.

In conclusion, The above Histograms depict the distribution of match wins across all the various venue where the concerned team has played the matches since the start of the league. The Bars depict the no. of wins at any particular venue while the curve is just a kernel density estimation for the data depicted by the bars in the plot providing for a smoother distribution metric.



Above is a pie chart with each sector of the chart representing teams that have been IPL champions, the percentage is calculated based on the no. of times the team has won in the Indian Premiere League.

For example:

No. of titles CSK has won = 5

Total no. of seasons played = 16

Share in the pie chart = (5/16)*100

= 31% (approx.)

The actual figure corresponds to the figure calculated during the analysis of the dataset.