**Predicting Trends and Analysis of IPL Data**

| **Student (s) Number as per your student card:** | **Rakshith Dheer Reddy Kunduru** | **20056561** |
| --- | --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |  |
| **Course Title:** | **MSc in Data Analytics** | |
|  |  | |
| **Lecturer Name:** | **Kanza Ali Manzar** | |
|  |  | |
| **Module/Subject Title:** | **Programming for Data Analysis** | |
|  |  | |
| **Assessment Title:** | **CA 2 - Data Analysis** | |
|  |  | |
|  |  | |

**ABSTRACT**

IPL in other words Indian Premier League is a world wide Tournament that is very popular nowadays. IPL is a cricket game that is played across india with preparing teams and choosing players under popular city names in india. We have around 19 teams give or take across the years. Each Team has around 15 players including main players that play and players that can substitute in need. Our goal is to analyse the patterns or trends from the IPL and how it can apply it to the real world. we can check in live if it applies to IPL 2025 that is currently ongoing

**PROBLEM STATEMENT**

The reason or motivation in choosing this dataset is that I have lost a certain amount of money in betting in IPL matches, so I was researching how I should predict as accurately as possible. As I was going through my course work of programming, ML and statistics which led to choosing IPL dataset over others for trend or pattern analysis for my CA 2 assignment.

**Preprocess and clean the dataset for Data Analysis**:

* Handle missing values, outliers, and inconsistencies.
* Normalize/standardize numeric features

**Feature engineering to extract meaningful insights:**

* Identify significant features and perform feature selection techniques.
* Explore correlations between features and the target.

**Conclusion:**

* Conclusion based on the plots that you have drawn.

**DATA DESCRIPTION**

The dataset used for this project is the "IPL" dataset, sourced from

**Kaggle:** [**https://www.kaggle.com/datasets/patrickb1912/ipl-complete-dataset-20082020**](https://www.kaggle.com/datasets/patrickb1912/ipl-complete-dataset-20082020)

This dataset contains IPL related matches data.

The dataset comprises two files. one file ‘matches.csv’ of **1095 rows** and **20 columns**. The Other File ‘delivieries.csv’ of **260920 rows** and **17 columns.** The first file records match id, season, city, date, match\_type, player of match, team1, team2, toss\_winner, toss\_decision, winner, result, result\_margin, target\_runs, target\_overs, super\_over, method, umpire1, umpire2.

The second file records match\_id, inning, battling\_team, bowling\_team, over, ball, batter, bowler, non-striker, batsmun\_run, extra\_runs, total\_runs, extras\_type, is\_wicket, player\_dismissal, dismassal\_kind and fielder.

**DATA PREPARATION**

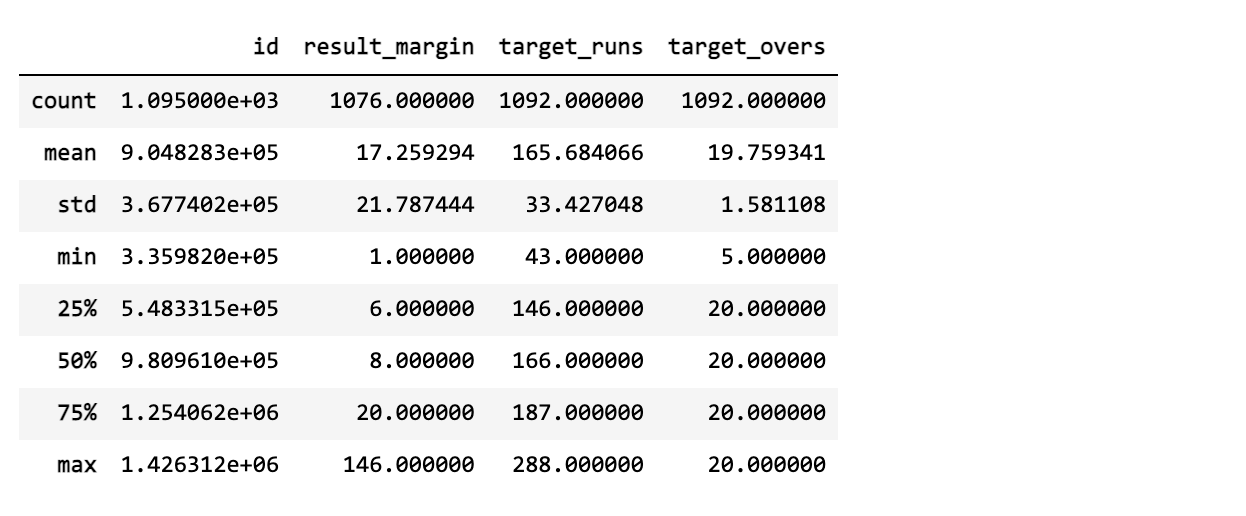
Data Preparation is a necessary process before model building or during the analysis of data.Data Preparation contains steps like Data Cleaning, Data Balancing, Data Encoding, Data Splitting and Data Scaling.

1. **Data Cleaning**

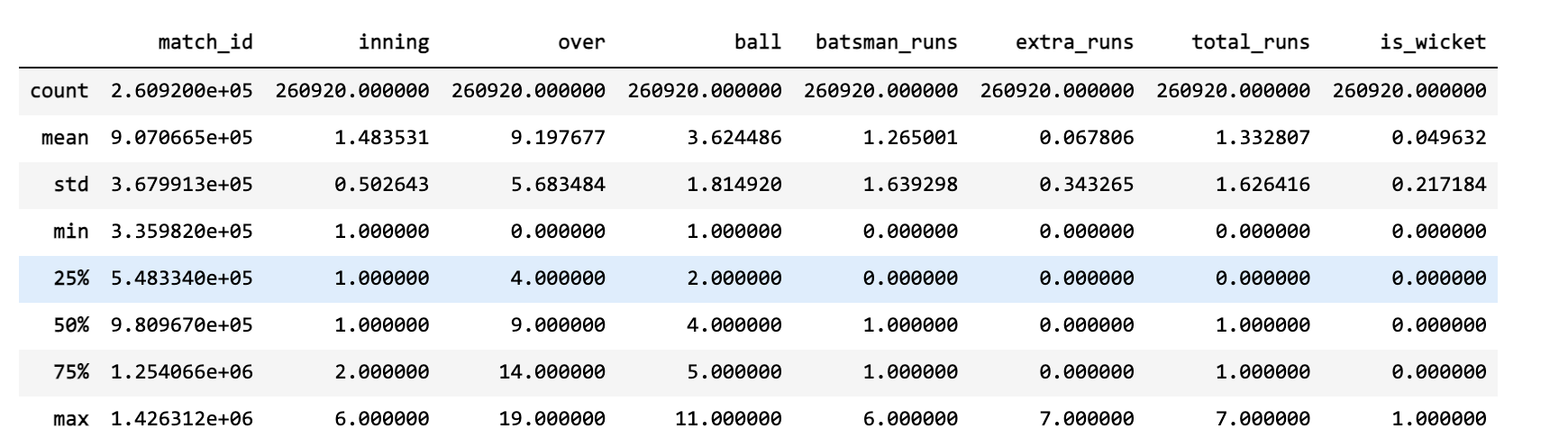
Data Cleaning is to remove unnecessary columns and missing values. By using few functions in pandas such as

data.info(),data.head(),data.describe()

Here is a an example for describe()function

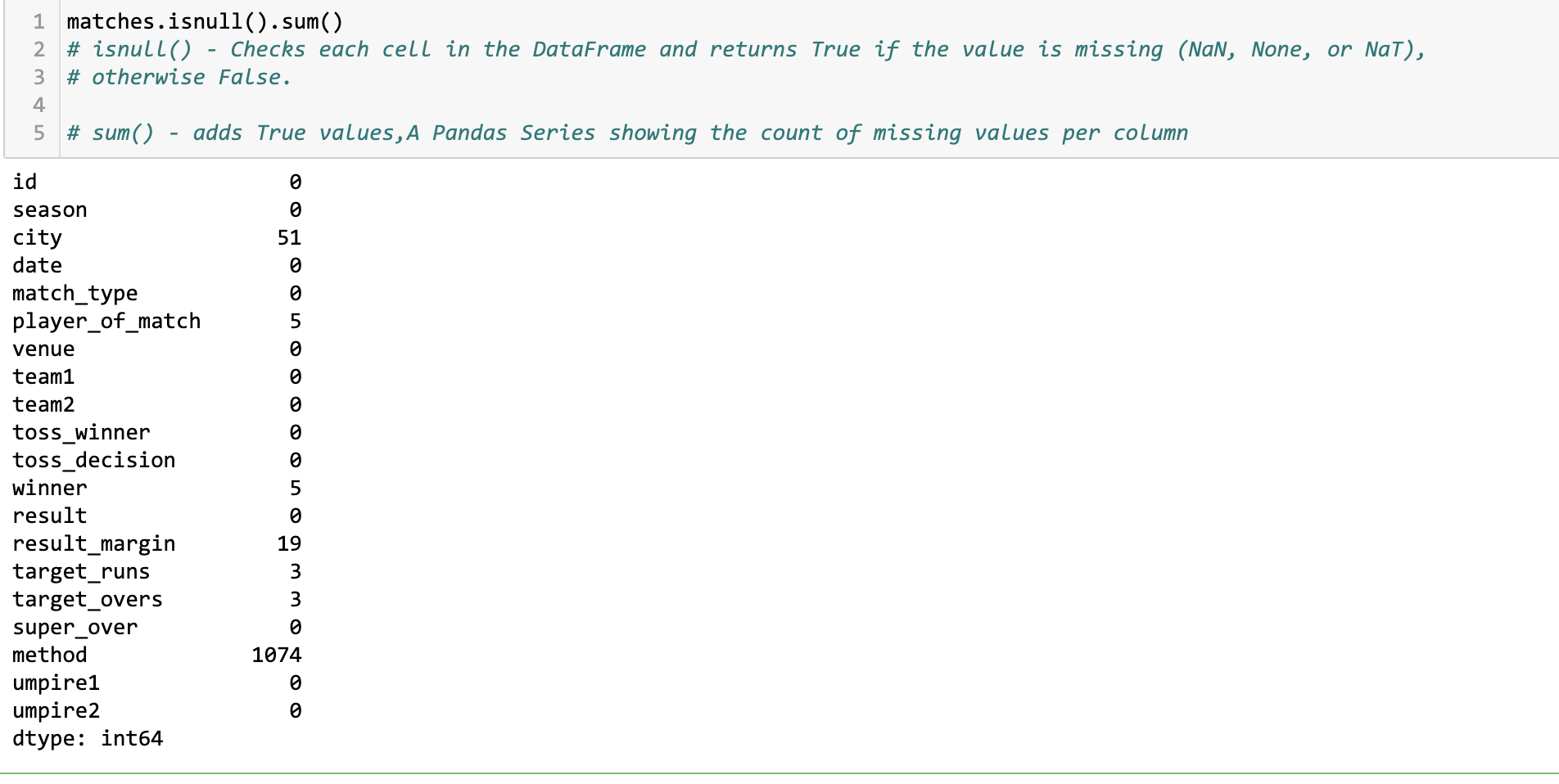


From this function output, we were able infer what is the mean of target\_runs and result\_margin.



We can find any outliers that are not in the normal range. Any missing values that are found can be removed or imputed with mean or median, if it is numerical or mode if it is categorical.

Here we can also decide to check rows which contain null values, using the below command and check the number of null values and take a call whether that can affect the data analysis or model building



1. **Data Encoding**

Statistics or model building behind the scenes is basically a calculation. Since the calculation requires numbers, we need to convert any text or string data type into numeric datatype or as per convenience. Since we cannot do math like *Yes \* 2.*

Here is an example,

we are mapping the seasons into string format, but we can also assign a season number as 1 using map functions



1. **Data Scaling**

We use Data Scaling to put all column units into a unit that is uniform among all the columns.

If we don’t use data scaling, the column which has a higher number due to its unit makes the other columns with lower number make its importance negligible.

According to dataset, it is like inputting strike rate, and runs into the data model

# Normalization/ Standardisation of the dataset

X\_scaled=StandardScaler().fit\_transform(X)

1. **Data Splitting**

We Split the data to train and test the model such that we can know if the model is deployable globally before it passes a certain business criteria.

Here is an sample Code

# Splitting the dataset into training and testing

X\_train,X\_test,Y\_train,Y\_test= train\_test\_split(X\_scaled,Y, test\_size=0.3,random\_state=1)

1. **Data Balancing**

Data Balancing is necessary to make the model that both the outcomes are of 50% chance. If we don’t do this step, the output model will be heavily biased to outcome with higher number of outcomes.

Here is an Sample Code

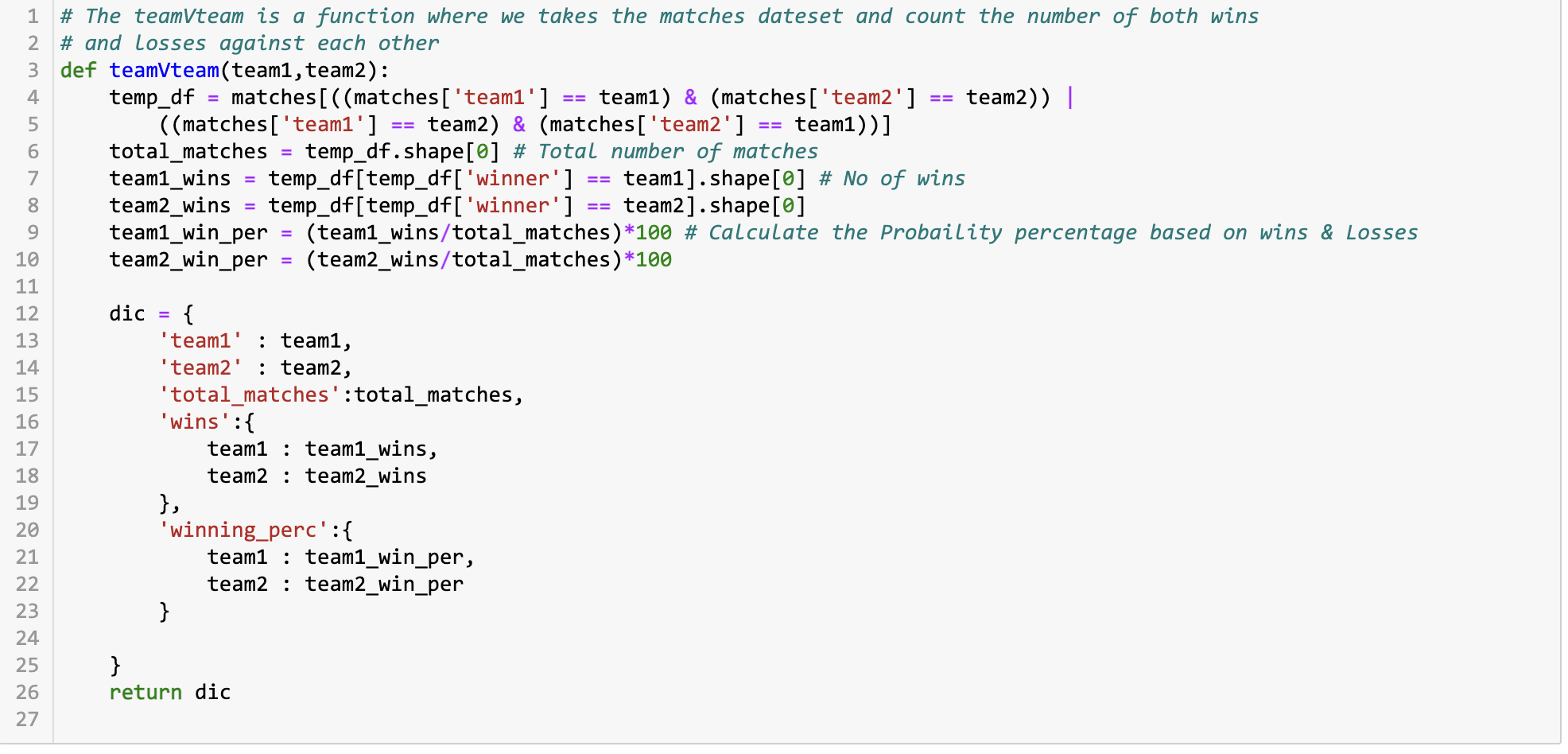
# Balancing the dataset using SMOTE technique

X\_train, Y\_train=SMOTE(random\_state=1).fit\_resample(X\_train,Y\_train) # Balancing the class labels

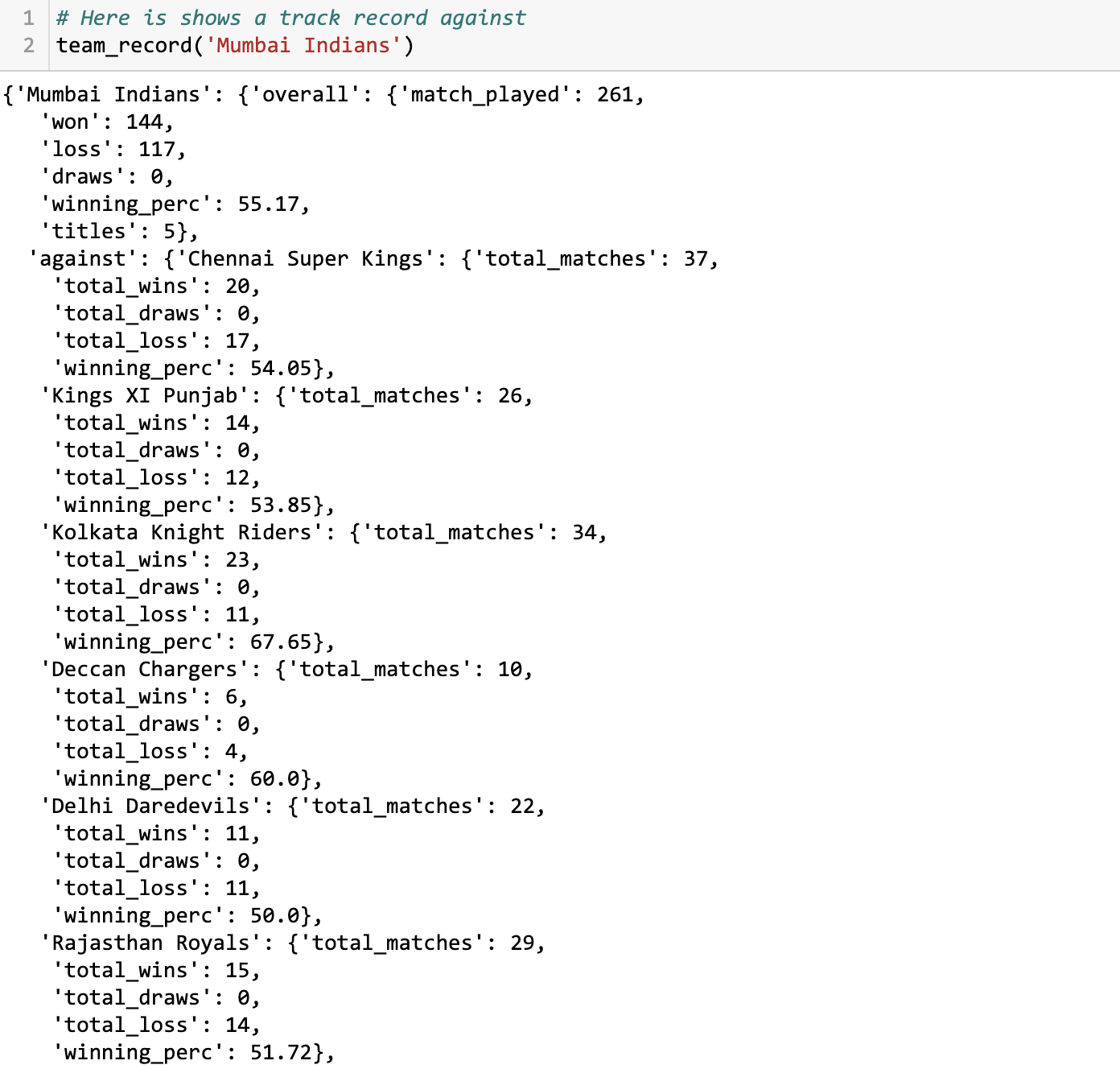
In this scenario, we are not yet prepared for model building, after considering and addressing the challenges faced we will proceed to Model Building in my future work.

**Exploratory Data Analysis - Find Patterns**

**01. Team Vs Team Analysis**

****

In the above image, we are comparing each team against another team. In the Code, we take the matches data file and count the given team wins, losses and draws against each other team through which we also calculate the winning percentage of the team when the same two teams are playing against each other.

****

In the above image, we have taken the example of team ‘Mumbai Indians’. It shows how Mumbai Indians fare against Chennai Super Kings, Kings X1 Punjab etc… We were able to infer roughly who has the higher chances of winning against other teams.

For Example, In this March 23rd 2025, for IPL 2025, mumbai indians and chennai super kings played against each other and CSK (Chennai Super Kings) won MI (Mumbai Indians) by 4 wickets. The analysis or inference clearly supports the future data.

**02. Batsman Vs Team Analysis**

The Batsman Vs Team is a very important factor for predicting the outcome.As the team members other than the core members change each year based on the auction.

So we have to consider this information or keep it in mind if there are any dangerous players who can change the outcome of the match by himself which may lead to loss of money.

****

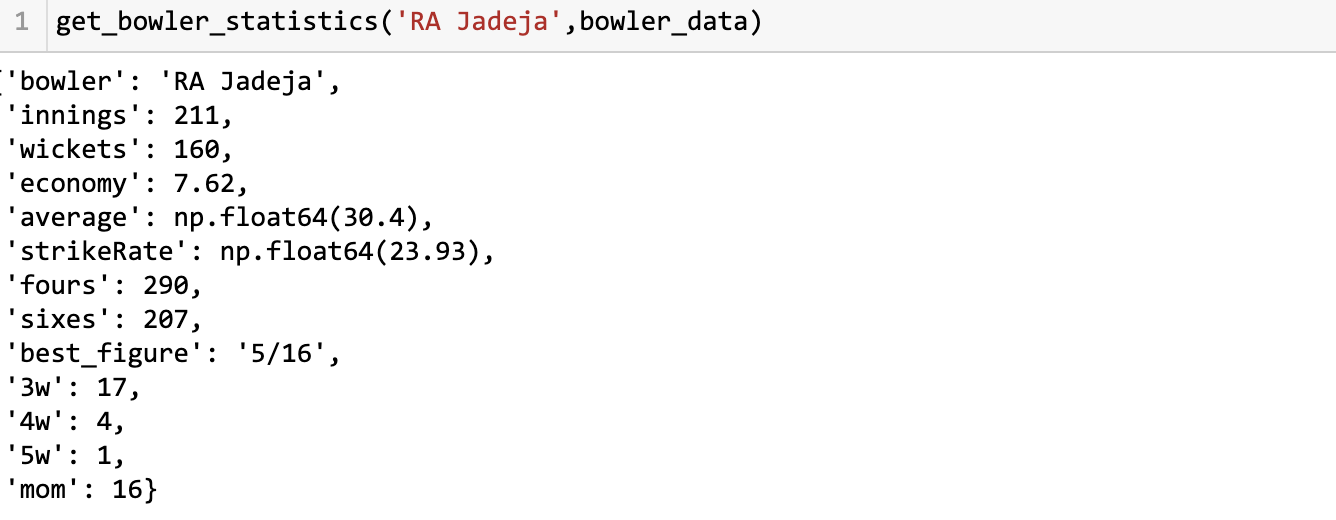
Here we have seen the statistics of one of the powerful batsmen and upcoming captain for the Indian national team who has this capability. We can also infer from this data whether the player is in the form or not and how well we can play in the upcoming innings.

*Challenge :* If the player is injured or for any other circumstances, the weight put in player for team winning like, “virat is present so he can chase this score” which the current data or analysis may not support it currently,

**03. Bowler Vs Team Analysis**

The Bowler Vs Team is a very important factor for changing the outcome in a surprising manner, A good bowler can change the outcome of the match

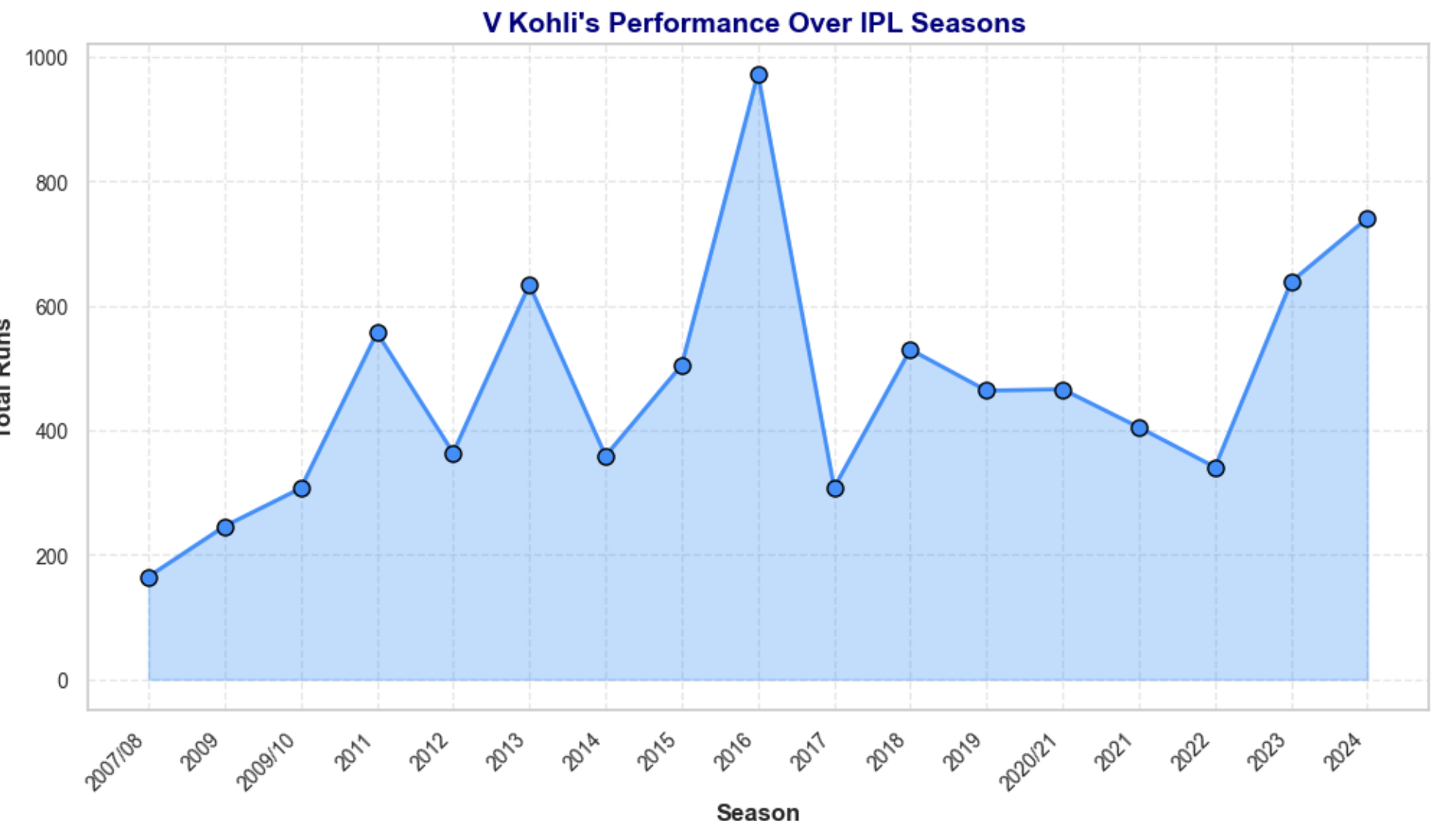
So we have to consider this information or keep it in mind if there are any dangerous players who can change the outcome of the match by himself which may lead to loss of money if you bet on the opposite team winning.

****

**04. Performance of Batsman over Seasons**

The Performance of Batsman over seasons help us identify who has the most consistent performers across seasons. As this cricket is a very volatile scenario, A single over can change the whole outcome of the match.

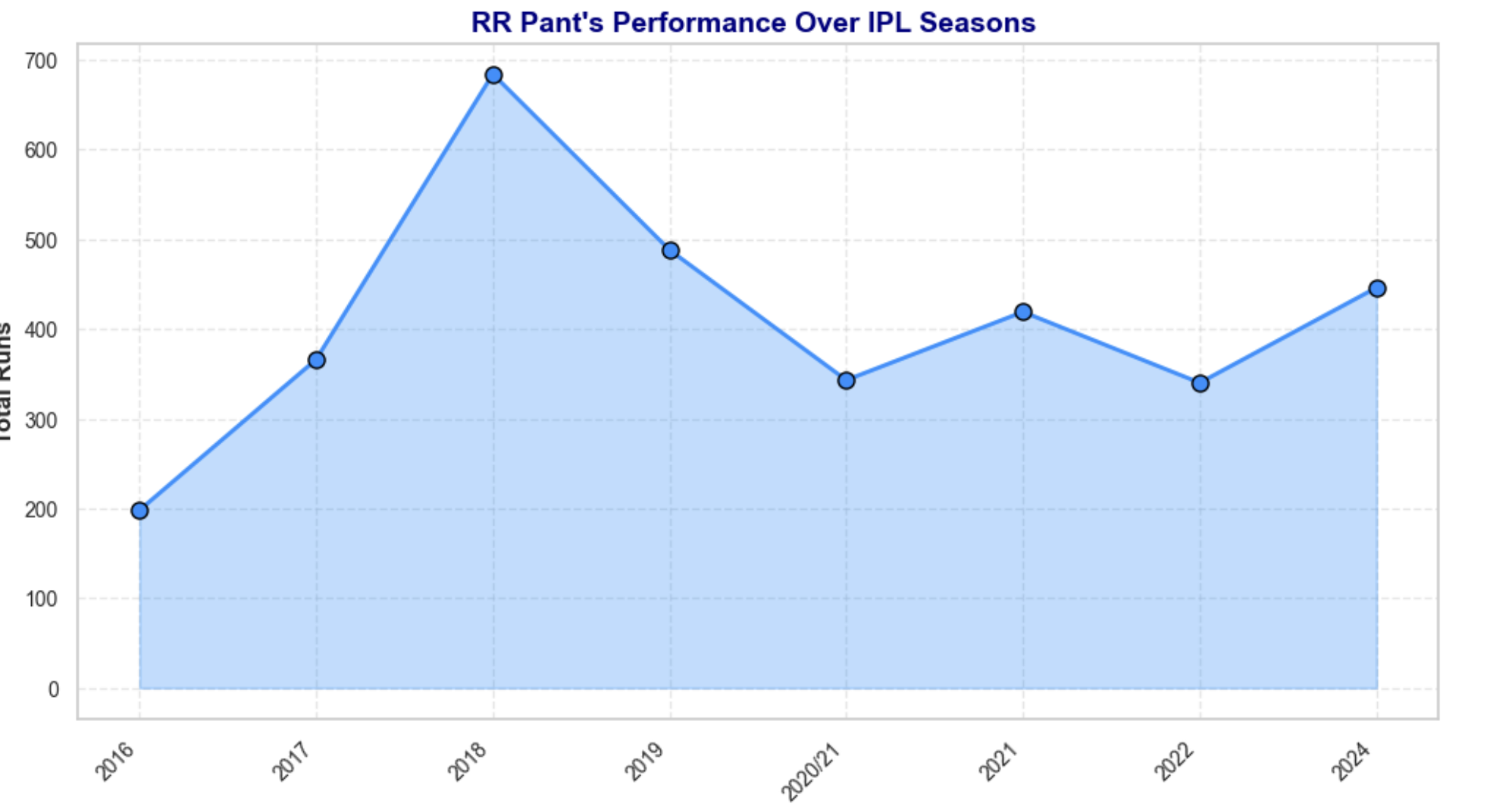
The Graph below shows on the super star of cricket virat kohli who has a spectacular performance across seasons which supports the increasing number of fans and weight towards him in BCCI (The Board of Control of Cricket in India)

****

This graph is also used by leaders or owners of the teams while selecting or buying the players for their own team. As there is a lot of money involved in buying the player. In the latest IPL 2025,

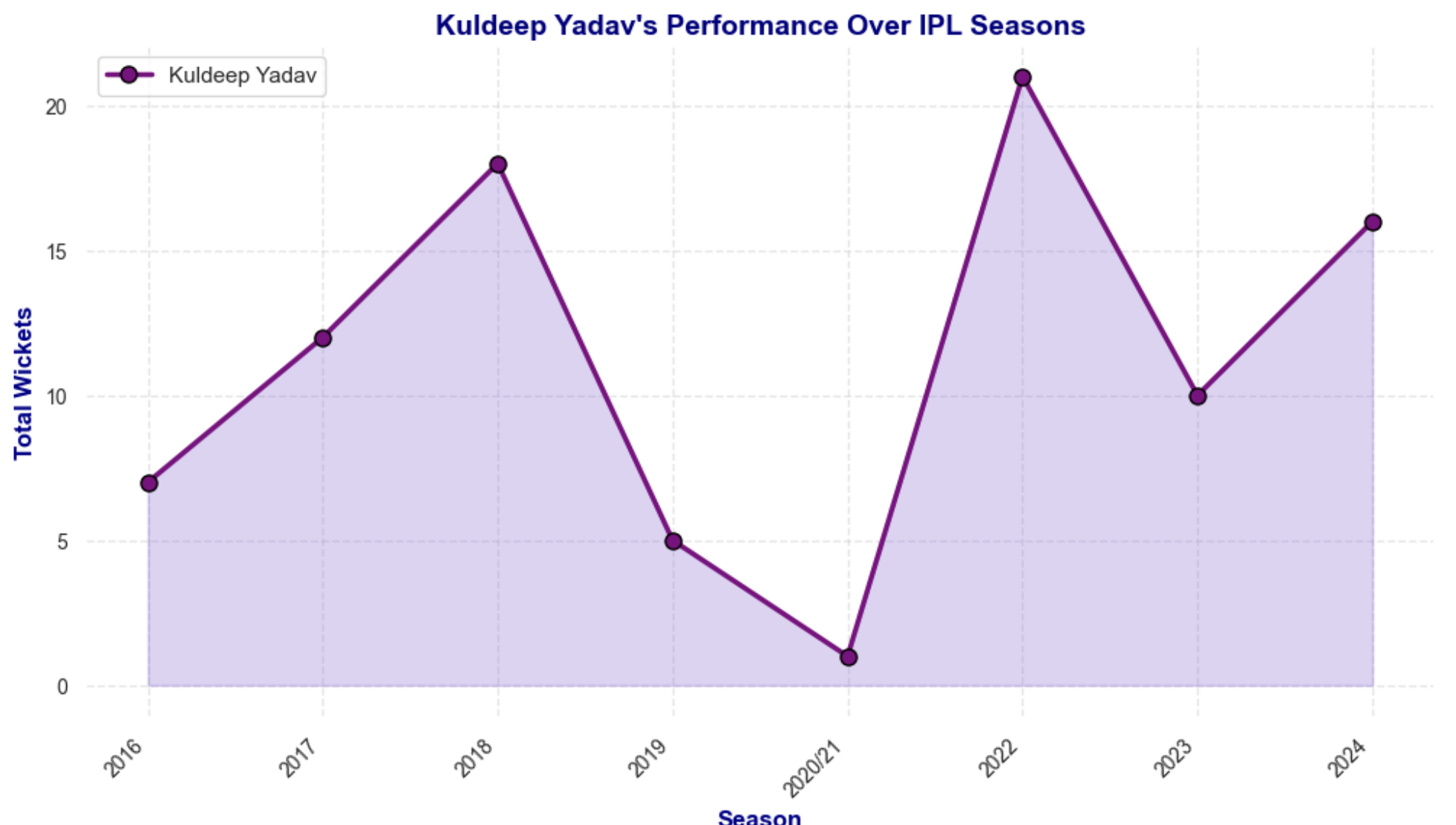
Lucknow SuperGiants has bought Rishabh Pant for 27 Crores. As he is one of the most outstanding players in the previous IPL seasons.

Below shows a graph of his performance across seasons and made him the highest bought player in IPL 2025.

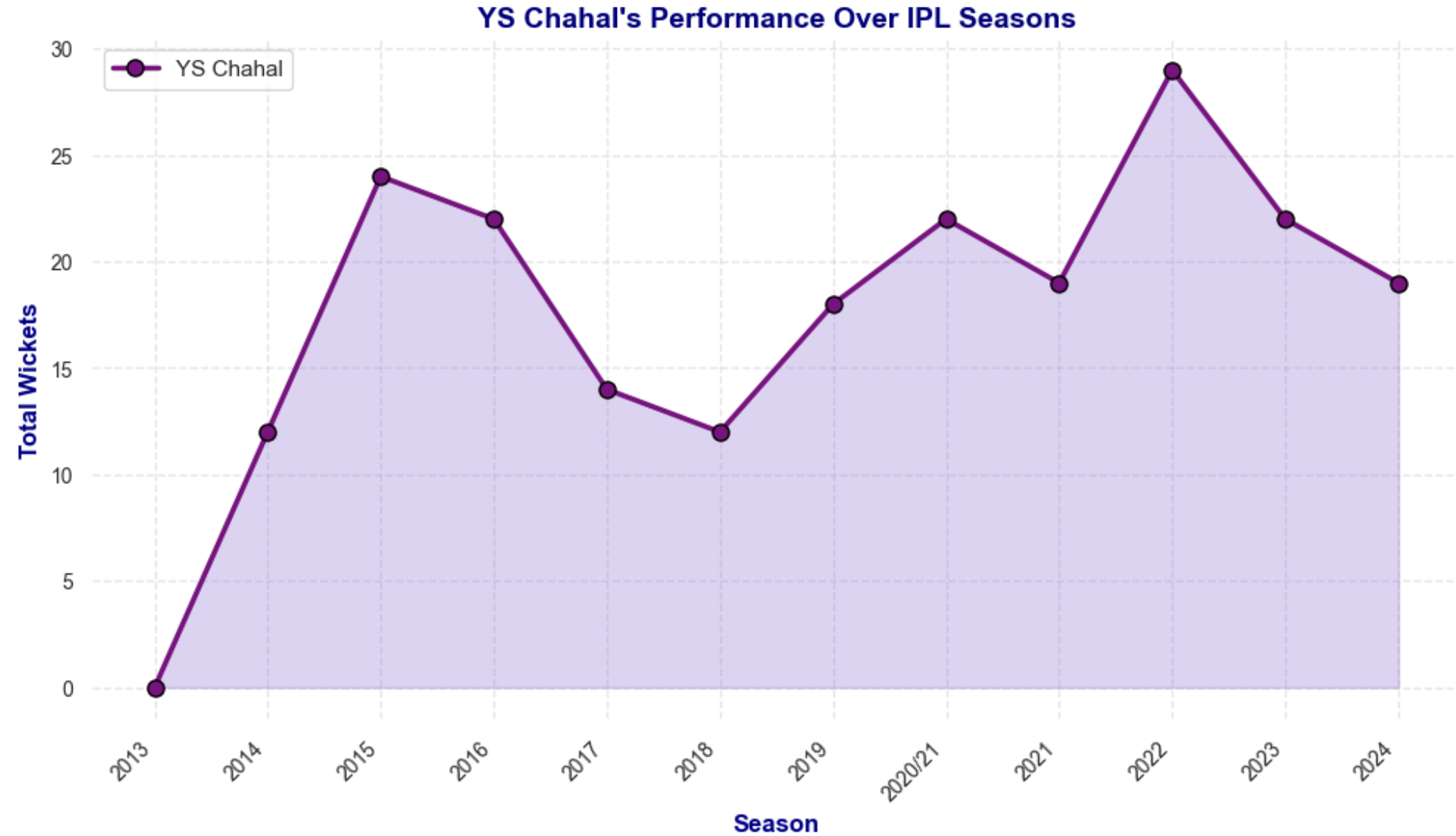


**05. Performance of Bowler over Seasons**

The Performance of Bowlers to take wickets can turn the match outcome in a surprising manner. The Number of Wickets a person takes in the season or a match plays a significant role in increasing his team ranking in the season.

****

The below plot shows the chahal who is the most expensive bowler in IPL 2025. As shown in the trend, the insights while using the performance is already being applied nowadays which leads me to thinking and imagine that data analysis plays a vital role in every industry.

****

**CHALLENGES**

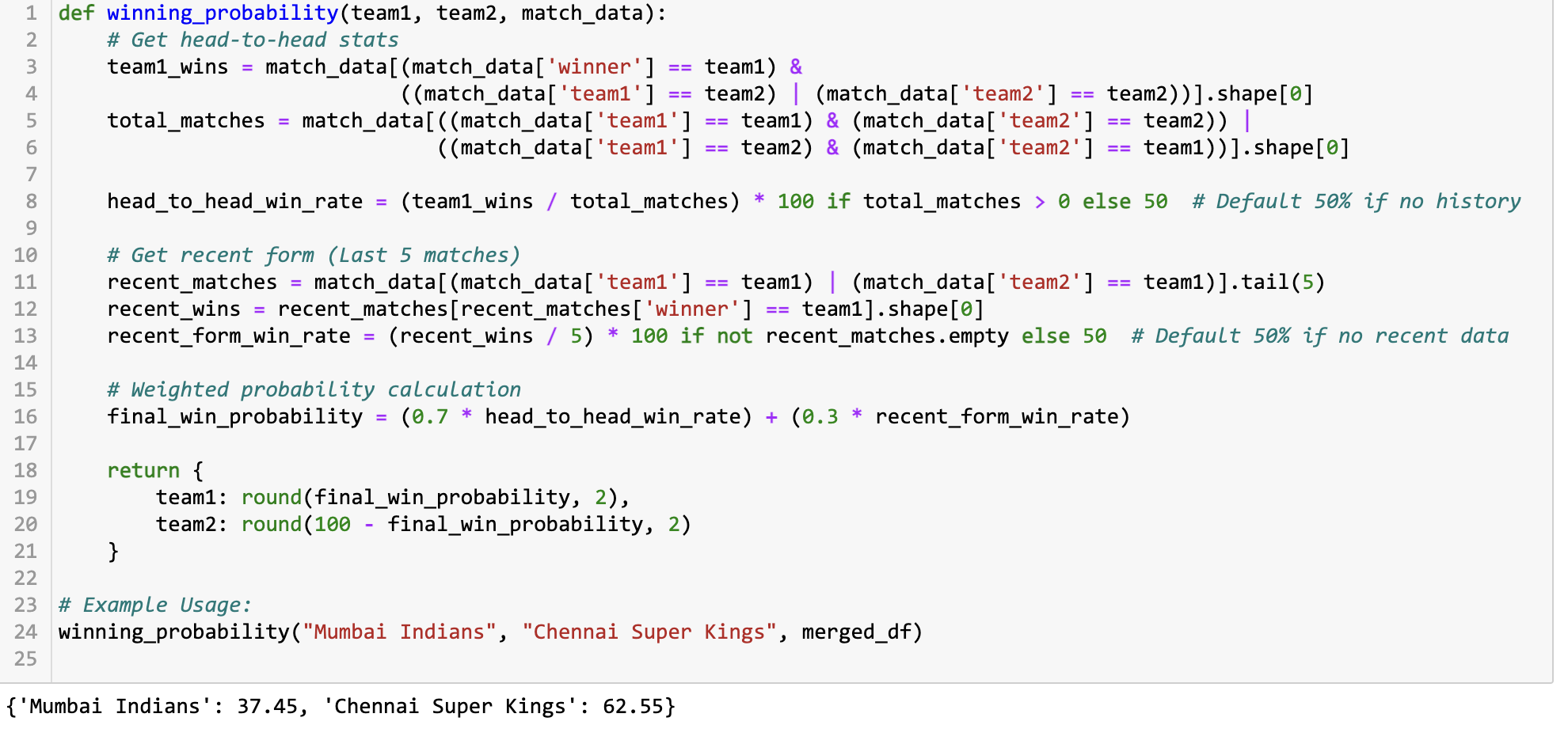
* Including latest condition or situation of the player like injuries into analysis
* Building a model that can take all the patterns, not just one pattern and predict the outcome. Here we have to consider above all five of them to predict a win or loss
* Analysing the score of new players, as there is no information currently for debuting players which may become a breaking point for the prediction outcome.

**FUTURE WORK**

* If there any players shift the team, based on the current team, the model or analysis should be able to predict the outcome
* All the while we are talking, predicting the winning the match, There can also be analysis for each batter against bowler and current fielders including over to be able to calculate runs should be able score per over. As there will be a over wise betting with very odds for each match
* Should be able to build a web app that shows the analysis and helps users bet on online platforms.

**CONCLUSION**

Here shows a code that can predict an outcome between the two teams and also gives the probability of winning percentage. It also considers not the team1 against team2 but also overall statistics of teams 2 and calculates probability which is better than just calculating the probability percentage of winning.

****

The Probability shows when Mumbai Indians and Chennai Super Kings go against each other, Chennai Super kings will win by 62%.

In the 2025 IPL, MI Vs CSK 38th Match was held in wankhede stadium on April 18th. Let’s Check whether the statistics wins or a miracle happens

**REFERENCES**

Briggs, P. (2021) \*IPL Complete Dataset (2008-2020)\* [Dataset]. Available at: https://www.kaggle.com/datasets/patrickb1912/ipl-complete-dataset-20082020 (Accessed: 10 July 2024).

McKinney, W. (2010) \*pandas: Powerful data structures for data analysis\*. Available at: https://pandas.pydata.org (Accessed: 15 July 2024).

Harris, C.R. et al. (2020) \*NumPy: Array computing for Python\* [Software]. Version 1.24.0. Available at: https://numpy.org (Accessed: Day Month Year).

Hunter, J.D. (2007) \*Matplotlib: A 2D graphics environment\* [Software]. Version 3.7.0. Available at: https://matplotlib.org (Accessed: Day Month Year).

Waskom, M.L. (2021) \*Seaborn: Statistical data visualization\* [Software]. Version 0.12.2. Available at: https://seaborn.pydata.org (Accessed: Day Month Year).