A picture containing icon

Description automatically generated­



ANALYSTS

Data Visualization and Exploratory Data Analysis of Indian Premier League

(2008 - 2020)

Anand Prakashbhai Shroff

Ravi Ravi

Santhosh Pachaiyappan

Simranjeet Kaur

Vignesh Dharuman

**TABLE OF CONTENTS**

1. [INTRODUCTION 3](#_TOC_250014)
   1. [Background 3](#_TOC_250013)
   2. [Problem Statement. 3](#_TOC_250012)
   3. [Scope 3](#_TOC_250011)
   4. [Document Overview 4](#_TOC_250010)
2. [PRELIMINARY REQUIREMENTS 4](#_TOC_250009)
3. [TECHNICAL APPROACH 4](#_TOC_250008)
   1. [Analysis 4](#_TOC_250007)
   2. [Requirements Development 5](#_TOC_250006)
   3. [Model Development 5](#_TOC_250005)
   4. [Testing and Evaluation 5](#_TOC_250004)
   5. [Delivery 5](#_TOC_250003)
4. [DESCRIPTIVE STATISTICS 4](#_TOC_250008)
   1. [Approach and Table descriptions 4](#_TOC_250007)
5. [EXPLORATORY DATA ANALYSIS](#_TOC_250008) 11
   1. Match [Analysis](#_TOC_250007) 12
   2. [Players Analysis](#_TOC_250006) 15
6. **DATA CLEANING AND TRANSFORMATIONS……………………………………………………. 18**
7. DATA ANALYSIS………………………………………………………………………………………….25
8. [EXPECTED RESULTS](#_TOC_250002) 33

9 REFERENCES 33

**1. INTRODUCTION**

**1.1 Background**

After India’s first world cup win in 2007 after 24 years, it was major tournament win for INDIA after a very long time. BCCI (Board of Cricket Control) took this opportunity as a golden chance to start a business of franchise based fast format cricket league which will be known as IPL (Indian Premier League) and the idea was similar to premier league of England and NBA of United states. The tournament has been a successful business venture and excellent instrument for cricket lovers from across the globe and have market value of US $6.3bil dollars in last 13 years which is an explosive growth.

The current dataset is a on which team FIVE ANALYST will be working is a clustered data which is separated in 8 excel data files and data consisting of Ball-by-ball performance players, team performances, Win: Lose rate of home vs away matches and player details.

There are final goals to achieve with the help of this project are as follows:

Develop game pre- and post-game analysis report for commentators, team stations, sports analysts, and cricket experts, which will help teams to plan their line-up before game and get better knowledge of terrain they are playing in.

Sportsmen and sportswomen have always been one of the prime choices for marketing any products or taking place of brand ambassadors. With the help this project’s predictive analysis companies will get better idea about selecting best players for their product campaigns.

There are many mobile apps such as Dream 11, MPL, Paytm first games and WINzo where players around the world win big cash prizes for their small investments if their players they select perform well in match. Cricket enthusiasts can find this visualization useful as a direction to put their investments on.

**1.2 Problem Statement**

It is crucial for every franchisee to analyse and evaluate past performances of their team so as to focus and improve on their weak and strong aspects. Our team will be performing a descriptive data analysis of IPL matches played between 2008-2020 in order to gain meaningful insights and recognize patterns and relationships between various attributes such as how many times team winning the toss won the match as well, how batting first or bowling first affects chances of winning, number of matches won at home ground and away, dependency on top order and middle order to perform well in batting, etc.

Scope

* This project has a vast amount of data to work with and develop some helpful insights based on different aspects, our project team will be working on the task with data which is vital to predict certain outcomes and will eventually be able to be visualized
* Constructing a model based on the players and the average runs they scored to predict the outcome of their performance in the upcoming game.
* Comparing the home wins and away wins of multiple teams to visualize which team is stronger on the home ground and which one is stronger on the away ground, and which are all keeping up in both, these visualizations will help figure out the strong winning ratios of teams based upon the ground they play at.
* It can be only taken into consideration of those data which can be visualized and provide wide insight into the task at hand.
* On visualizing performance of the player on the venue also helps to predict the runs the player will score in same condition
* Comparing bowler vs batsman also helps to identify the winning probability of the match

**1.3 Document Overview**

The dataset which is used for project is IPL data from year 2008-2020. The main requirements and results of project are sub-divided into various categories which are explained in this proposal under name of Preliminary requirements, technical approach, expected results and management approach. Each section explains how our team is going to use step by step approach to complete this project by describing problem statement and achieve the scope of this project.

**2. PRELIMINARY REQUIREMENTS**

**Data Requirement:**

The complete data of India premier league from 2008 to 2020, it should have Bowling stats, batting stats, Season details, ball by ball match information, venue details etc.

The dataset is collected from below links

[**https://www.kaggle.com/ramjidoolla/ipl-data-set**](https://www.kaggle.com/ramjidoolla/ipl-data-set)

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

[**https://www.kaggle.com/datasets/rajsengo/indian-premier-league-ipl-all-seasons?select=all\_season\_batting\_card.csv**](https://www.kaggle.com/datasets/rajsengo/indian-premier-league-ipl-all-seasons?select=all_season_batting_card.csv)

[**https://www.kaggle.com/datasets/rajsengo/indian-premier-league-ipl-all-seasons?select=all\_season\_bowling\_card.csv**](https://www.kaggle.com/datasets/rajsengo/indian-premier-league-ipl-all-seasons?select=all_season_bowling_card.csv)

**Business Requirement**

|  |  |
| --- | --- |
| **S.no** | **Functional requirement** |
| 1 | Visualize the data in graphs of each season winners, toss winners and total number of wins |
| 2 | Geographical visualization of venues, player progress, team progress and history of venues |
| 3 | Plot graphs, bar chart, bubble chart, pie chart of all the data provided |
| 4 | Pictorial representation of player stats from top to least |
| 5 | Represent dashboard with tabs such as Team, venue, Player, and seasons |
| 6 | Home dashboard with all key details and links to details information that navigates to more stats |
| 7 | Complete dashboard in tableau which should look like a website |

**3.TECHNICAL APPROACH**

**3.1 Analysis**

In order to carry out this analysis successfully team has to understand the following points:

Learn to perform the ETL (Extract Transform Load) basics before starting to analyze data, how to implement the variables in the visualization.

Learn about the relation between all csv sheets and how to join key variables across the dataset, which variable to choose for dataset displaying and which to ignore.

Team also has to learn software like tableau and excel analysis and how to use them to draw visualizations from the clustered dataset.

**3.2 Requirements Development**

Knowledge of Microsoft Office tools which will be used for understanding the data and its attributes thoroughly. It will help in scoping and identifying potential relationships to be explored within the dataset.

Understanding and working of MS Excel and Python for data cleaning, preparation, and visualization.

Working with Tableau, where the dashboard will be developed based on the project

**3.3 Model Development**

The first thing to consider in the development of the model is to take the inputs which are needed, which can be achieved by analysing the complete data set and by taking all the aspects into consideration. The player average and the winning ratios of the teams will be visualized by patterns of graphs.

Further visualizations will be done by sorting out the data in Excel by combing through each and every aspect of individual rows and columns and then with the use of python which will be manipulating the data to figure out more useful insights as it gets to the point where the visualization can be done, Tableau will be used to do the final visualization.

**3.4 Testing and Evaluation**

An evaluation of whether the data model visualize the correct information and in right manner is dashboard. Testing and evaluation of product is categorized into multiple process such as Cosmetics bugs which includes spelling, grammatical and presentation. And data testing, plotting and graphs also tested and evaluated. Input data and output result also evaluated in each step.

Testing and evaluation can be split into multiple phases as follows

* **Unit Testing:**

In this phase, data is tested, and input and output of each plotting is evaluated. Testing methods used are Component testing, UI testing and Sanity testing

* **Integration Testing:**

In this phase, Plots and graphs are integrated one by one and evaluated. Testing methods used are integration testing, Regression testing, Smoke testing and Sanity testing

* **System Testing:**

In this phase, Plots and graphs are integrated into one dashboard in local server and evaluated. Testing methods used are integration testing, Regression testing, Smoke testing and Sanity testing, compatibility test, load test, Performance test.

* **User acceptance testing:**

In this phase, the pilot product is prepared and is hosted online and evaluated as user. Testing methods used are Regression testing, Sanity testing, load test, Performance test and connectivity testing and UAT testing.

**3.5 Delivery**

Final delivery will be done as a tableau dashboard which will be the in the form of dashboard charts which will clearly depict the conclusions drawn by the dataset. The results will have information about relation of Ground conditions with player performances, win ratio, player stats, player growth and will also work on match win prediction, and player performance stats in relation the terrain (ground) or certain bowler types.

**4. DESCRIPTIVE STATISTICS**

**4.1 Approach and Data descriptions**

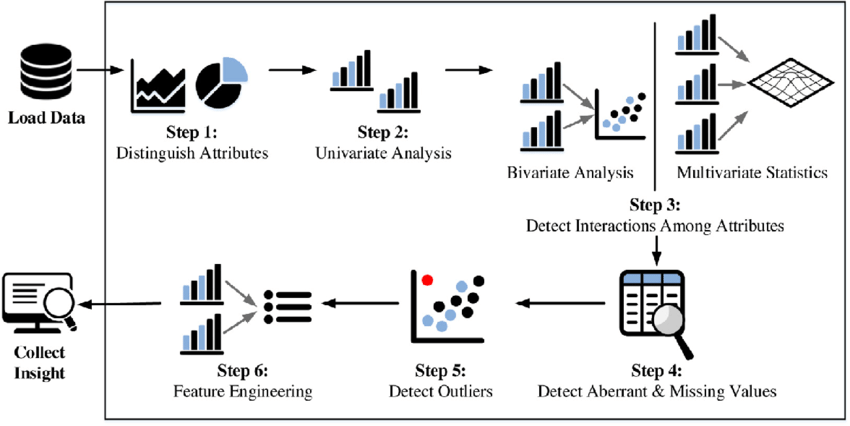


Figure 1. Steps followed in Descriptive statistics

The above steps are followed in identifying the dataset details which are categorized in match stats and player stats

Diagram

Description automatically generated

Figure 2. Variables split up

Table 1. IPL Match (2008-2020) Data

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Column name | Data Type | Count | Frequency (Unique) | Null values | NA values | Variable type |
| id | Int64.Type | 816 | 816 | 0 | 0 | Continuous |
| city | type text | 816 | 33 | 0 | 13 | Categorical |
| date | type date | 816 | 596 | 0 | 0 | Continuous |
| player\_of\_match | type text | 816 | 234 | 0 | 4 | Categorical |
| venue | type text | 816 | 36 | 0 | 0 | Categorical |
| neutral\_venue | Int64.Type | 816 | 2 | 0 | 0 | Categorical |
| team1 | type text | 816 | 15 | 0 | 0 | Categorical |
| team2 | type text | 816 | 15 | 0 | 0 | Categorical |
| toss\_winner | type text | 816 | 15 | 0 | 0 | Categorical |
| toss\_decision | type text | 816 | 2 | 0 | 0 | Categorical |
| winner | type text | 816 | 16 | 0 | 4 | Categorical |
| result | type text | 816 | 4 | 0 | 4 | Categorical |
| result\_margin | type text | 816 | 92 | 0 | 17 | Categorical |
| eliminator | type text | 816 | 3 | 0 | 4 | Categorical |
| method | type text | 816 | 2 | 0 | 797 | Categorical |
| umpire1 | type text | 816 | 48 | 0 | 0 | Categorical |
| umpire2 | type text | 816 | 47 | 0 | 0 | Categorical |

Table 1 describes the complete details for matches from 2008 to 2020. It has 816 rows and 17 columns. Major columns which are used here is winners, toss winners, player of match, toss decision, and venue.

Table 2. Teamwise\_home\_and\_away.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Column name | Data Type | Count | Frequency (Unique) | Null values | NA values | Variable type |
| team | type text | 14 | 14 | 0 | 0 | Categorical |
| home\_wins | Int64.Type | 14 | 14 | 0 | 0 | Continuous |
| away\_wins | Int64.Type | 14 | 13 | 0 | 0 | Continuous |
| home\_matches | Int64.Type | 14 | 14 | 0 | 0 | Continuous |
| away\_matches | Int64.Type | 14 | 13 | 0 | 0 | Continuous |
| home\_win\_percentage | type number | 14 | 14 | 0 | 0 | Continuous |
| away\_win\_percentage | type number | 14 | 14 | 0 | 0 | Continuous |

Table 2 describes the details for home winner’s vs away winners from 2008 to 2020. It has 14 rows and 7 columns. It has one categorical and 6 continuous numerical variables

Table 3. IPL Ball-by-Ball 2008-2020

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Column name | Datatype | Count | Frequency (Unique) | Null Values | NA Values | Variable type |
| id | Int64.Type | 193468 | 816 | 0 | 0 | continuous |
| inning | Int64.Type | 193468 | 2 | 0 | 0 | Numerical |
| over | Int64.Type | 193468 | 20 | 0 | 0 | Numerical |
| ball | Int64.Type | 193468 | 9 | 0 | 0 | continuous |
| batsman | Type.Text | 193468 | 537 | 0 | 0 | categorical |
| non\_striker | Type.Text | 193468 | 530 | 0 | 0 | categorical |
| bowler | Type.Text | 193468 | 420 | 0 | 0 | categorical |
| batsmans\_runs | Int64.Type | 193468 | 7 | 0 | 0 | continuous |
| extra\_runs | Int64.Type | 193468 | 7 | 0 | 0 | continuous |
| total\_runs | Int64.Type | 193468 | 8 | 0 | 0 | continuous |
| non\_boundary | Int64.Type | 193468 | 2 | 0 | 0 | continuous |
| is\_wicket | Int64.Type | 193468 | 2 | 0 | 0 | categorical |
| dismissal\_kind | Type.Text | 193468 | 10 | 0 | 183973 | categorical |
| player\_dismissed | Type.Text | 193468 | 507 | 0 | 183973 | categorical |
| fielder | Type.Text | 193468 | 880 | 0 | 186684 | categorical |
| extras\_type | Type.Text | 193468 | 6 | 0 | 183235 | categorical |
| batting\_team | Type.Text | 193468 | 15 | 0 | 0 | categorical |
| bowling\_team | Type.Text | 193468 | 16 | 0 | 191 | categorical |

Table 3 describes the complete ball by ball details of matches from 2008 to 2020. It has 193468 rows and 18 columns. Major columns which are used here is wicket, runs, extras type, batsman details and bowler details.

NA values in IPL Ball-by-Ball 2008-2020 Table doesn’t means it is not available, it actually linked with this column “dismissal\_kind, player\_dismissed, fielder, extras\_type”. For example, if one column has value other 3 columns will be NA.

Table 4. Player details

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Column name | Data Type | Count | Frequency (Unique) | Null values | NA values | Variable type |
| Player\_Name | Text | 566 | 566 | 0 | 0 | Categorical |
| DOB | Date | 471 | 457 | 95 | 0 | Continuous |
| Batting\_Hand | Text | 563 | 2 | 3 | 0 | Categorical |
| Bowling\_Skill | Text | 502 | 14 | 64 | 0 | Categorical |
| Country | Text | 471 | 11 | 95 | 0 | Categorical |

Table 4 describes the details for Players with name and their skills and country. It has 566 rows and 5 columns. It has one Continuous and 6 categorical variables

Table 5. All Season Bowling

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Column Name | Data Type | Count | Frequency | Null Values | NA values | Variable type |
| Season | Int64.Type | 9967 | 13 | 0 | 0 | Numeric(Int) |
| Match\_id | Int64.Type | 9967 | 816 | 0 | 0 | Numeric(Int) |
| Match\_name | type text | 9967 | 111 | 0 | 0 | Categorical |
| Home\_Team | type date | 9967 | 12 | 0 | 0 | Categorical |
| Away\_Team | type text | 9967 | 12 | 0 | 0 | Categorical |
| Bowling\_team | type text | 9967 | 12 | 0 | 0 | Categorical |
| Venue | type text | 9967 | 12 | 0 | 0 | Categorical |
| city | type text | 9967 | 12 | 0 | 0 | Categorical |
| country | type text | 9967 | 33 | 0 | 0 | Categorical |
| innings\_id | Int64.Type | 9967 | 12 | 0 | 0 | Categorical |
| name | type text | 9967 | 33 | 0 | 0 | Categorical |
| bowling\_style | type text | 9967 | 31 | 0 | 0 | Categorical |
| bowler\_country | type text | 9427 | 33 | 0 | 0 | Categorical |
| full\_name | type text | 9967 | 31 | 0 | 0 | Categorical |
| overs | Int64.Type | 9967 | 3 | 0 | 0 | Numeric (Float) |
| maidens | Int64.Type | 9967 | 31 | 0 | 0 | Numeric (Int) |
| conceded | Int64.Type | 9967 | 3 | 0 | 0 | Numeric (Int) |
| wickets | Int64.Type | 9967 | 2 | 0 | 0 | Numeric (Int) |
| economyRate | Int64.Type | 9966 | 3 | 1 | 0 | Numeric (Float) |
| dots | Int64.Type | 9967 | 2 | 0 | 0 | Numeric (Int) |
| fours conceded | Int64.Type | 9967 | 419 | 0 | 0 | Numeric (Int) |
| sixes conceded | Int64.Type | 9967 | 2 | 0 | 0 | Numeric (Int) |
| wides | Int64.Type | 9967 | 419 | 0 | 0 | Numeric (Int) |
| no balls | Int64.Type | 9967 | 13 | 0 | 0 | Numeric (Int) |
| captian | Type Boolean | 9967 | 419 | 0 | 0 | Boolean |
| herf | type text | 9967 | 13 | 0 | 0 | String |

Table 5 describes the details for bowling with name and their skills and country. It has 9967 rows and 25 columns. It has 14 Continuous and 11 categorical variables.

Table 6. most\_runs\_average\_strikerate

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Column Name** | **Data Type** | **Count** | **Frequency (Unique)** | **Null values** | **N/A values** | **Variable type** |
| Batsman | Text | 516 | 516 | 0 | 0 | CATEGORICAL |
| Total\_runs | int | 516 | 283 | 0 | 0 | CONTINUOUS |
| OUT | INT | 516 | 89 | 0 | 0 | CONTINUOUS |
| numberofballs | int | 516 | 266 | 0 | 0 | CONTINUOUS |
| average | int | 482 | 323 | 34 | 0 | CONTINUOUS |
| strikerate | int | 516 | 405 | 0 | 0 | CONTINUOUS |

Table 6 describes the details for batsman with name and their skills and country. It has 516 rows and 5 columns. It has one 5 Continuous and 1 categorical variable.

Table 7. all\_season\_batting\_card

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Column name** | **Data Type** | **Count** | **Frequency (Unique)** | **Null values** | **NA values** | **Variable type** |
| season | int | 12429 | 13 | 2 | 0 | CONTINUOUS |
| match\_id | int | 12429 | 816 | 2 | 0 | CONTINUOUS |
| match\_name | text | 12431 | 111 | 2 | 0 | CATEGORICAL |
| home\_team | text | 12431 | 12 | 2 | 0 | CATEGORICAL |
| away\_team | text | 12431 | 12 | 2 | 0 | CATEGORICAL |
| venue | text | 12431 | 33 | 2 | 0 | CATEGORICAL |
| city | text | 12431 | 31 | 2 | 0 | CATEGORICAL |
| country | text | 12431 | 3 | 2 | 0 | CATEGORICAL |
| current\_innings | text | 12431 | 12 | 2 | 0 | CATEGORICAL |
| innings\_id | int | 12429 | 2 | 2 | 0 | CONTINUOUS |
| name | text | 12431 | 541 | 2 | 0 | CATEGORICAL |
| fullname | text | 12431 | 541 | 2 | 0 | CATEGORICAL |
| runs | int | 12423 | 127 | 8 | 0 | CONTINUOUS |
| ballsfaced | int | 12423 | 70 | 8 | 0 | CONTINUOUS |
| minutes |  | 12431 | 111 | 8 | 0 | CONTINUOUS |
| fours | int | 12423 | 18 | 8 | 0 | CONTINUOUS |
| sixes | int | 12423 | 15 | 8 | 0 | CONTINUOUS |
| strikerate | int | 12431 | 1191 | 8 | 0 | CONTINUOUS |
| captain | boolean | 12429 | 2 | 2 | 0 | BOOLEAN |
| isNotOut | boolean | 12423 | 2 | 8 | 0 | BOOLEAN |
| runningScore | int, text | 12431 | 1522 | 8 | 0 | CONTINUOUS, CATEGORICAL |
| runningOver | int | 9484 | 125 | 2947 | 0 | CONTINUOUS |
| shorttext | text | 12431 | 4912 | 2 | 0 | CATEGORICAL |
| commentary | text | 12431 |  | 2950 | 0 | CATEGORICAL |
| total\_boundaries | int | 12431 | 25 | 0 | 0 | CONTINUOUS |

Table 5 describes the details for bowling with name and their skills and country. It has 12429 rows and 25 columns. It has 14 Continuous and 11 categorical variables.

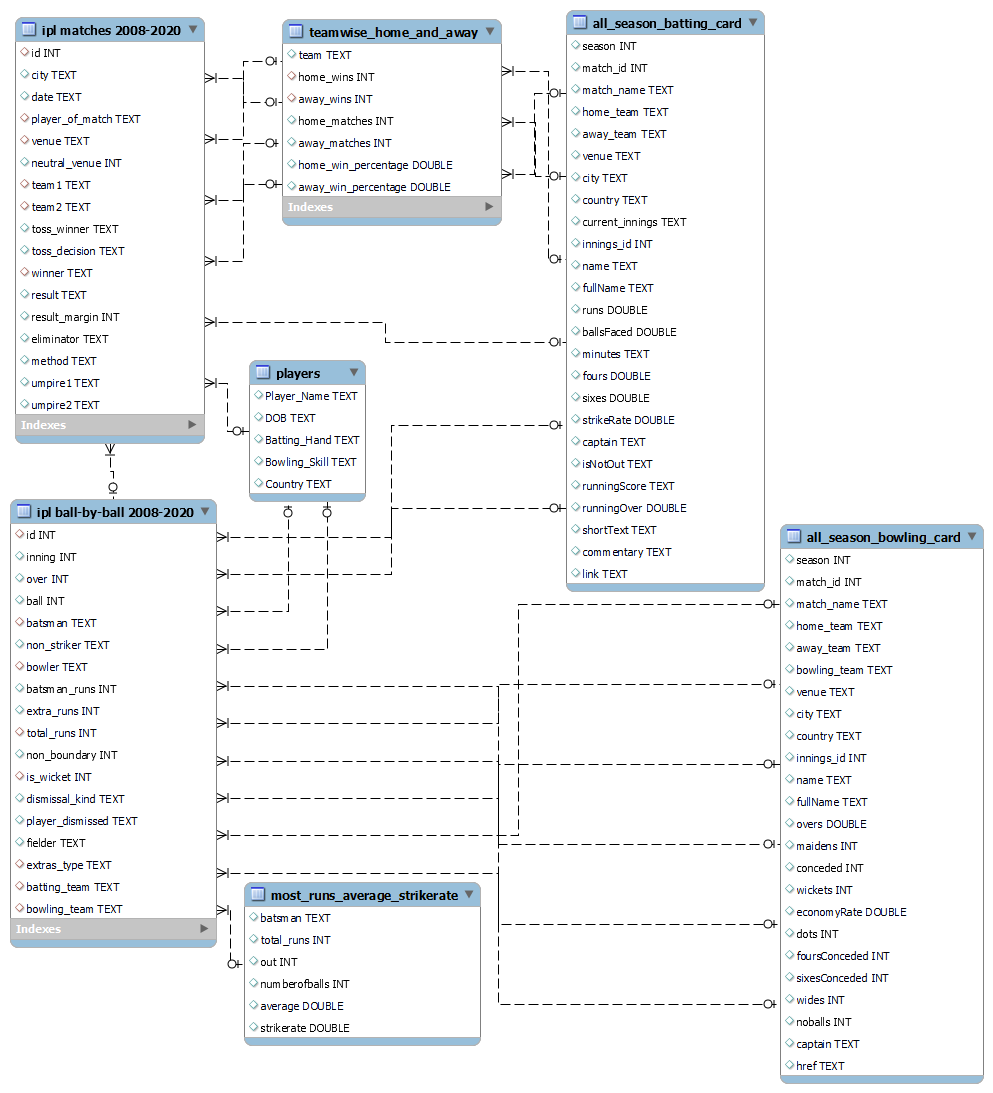


Figure 3. SCHEMA of IPL Dataset

**Outliers**

The IPL is mostly focused on Indian player, so the participants for Indian players will be more compared to foreign players. so, outliers related to players is not considered in this dataset. And the Venues are more concentrated in India though some matches are conducted out of India therefore outliers related to venue and place are still valid.

*Note: NA values in IPL Ball-by-Ball 2008-2020 Table doesn’t means it is not available, it actually linked with this column “dismissal\_kind,* *player\_dismissed, fielder, extras\_type”. For example, if one column has value other 3 columns will be NA.*

**5. EXPLORATORY DATA ANALYSIS**

* 1. **Match Analysis**

**Match and Toss Winners**

IPL is conducted every year where top 8 teams are selected for each season and each team will have minimum 14 matches in a season. Below chart show the number of matches won by each team in all seasons. In this it clearly shows that Mumbai Indian won highest number of match where the Kochi Tuskers Kerala. It also states that top 5 teams are participating in all seasons.

Figure 4. Match Winners

Figure 5. Toss Winners

The above chart describes the number of times each team won toss in each season. Mumbai Indians is the highest and followed by Chennai super kings and Kochi tuskers Kerala are the least. On comparing both chart it clearly shows that team who wins the toss has the higher chances of winning the match.

**Comparison of home and away wins**

The below table and chart describe the number of each team won in home and away ground(venue). Mumbai Indians has good records in home ground whereas Kolkata Knight Riders has good record in Away grounds. On comparing both home and away ground, Mumbai Indians has played a greater number of matches in home ground than the other teams.

Figure 6. Home Matches vs Home wins

Figure 7. Away Matches vs Away Wins

**Over analysis**

This data contains all the details of the balls bowled in both innings which also includes the bowler, batsmen, non-striker, fielder, wicket, and boundary details. This sheet also includes columns with NA values which are Player dismissed, dismissal kind, fielder and extras type, the reason behind that is all those data are depending on the outcome of the certain other column for example there will not be any player dismissed if there is no wicket in that ball.

Figure 8. Top 10 Bowlers with Most Extra Runs

This above chart depicts the data of the top 10 bowlers who has given most extra runs in the period of 2008 to 2020, S.L Malinga is the number one player who has given most runs which almost contributes to 34% and this chart can give an idea about how to deploy which player on which over which could be crucial at times when there should be no extra runs to control the unnecessary runs.

Figure 9. Top player's Runs based on overs

This above chart contains detailed data of total runs scored per over for both the innings from which it can identify the innings in which most of the runs is scored and also it will give a clear idea of runs scored in the overs to analyse which overs should be maintained to keep the overall scored leveled and this stat clearly shows which player has good stats based on the over. David warner has highest run percentage in first 5 overs and MS Dhoni has great run percentage in last 5 overs.

* 1. **Players Analysis**

**Batting**

**1. Number of players per Country**

IPL (Indian premier league) is a cricket league hosted by India but players from all countries participate in it under different team names. The chart below shows number of players from different countries which participated in IPL from 2008-2019. From chart it is clear that India has the greatest number of players (264) followed by Australia (72) and South Africa (39).

Figure 10. Number Of Players per Country

**2. Count of Players per country distributed according to batting style**

Figure 11. Country wise Batting style of Players

This chart shows the players distribution per country based on their batting. It’s clear from chart that Australia has most right-hand players after India. In charts it is shown that Netherlands and Zimbabwe have most right-hand players but count of players from both countries is 1 and 2 respectively which is less than Australia.

**3. Number of Boundaries per team**

Below graph shows the boundaries of each team which include Fours and sixes of each team.

Figure 12. Number of Boundaries per Team

This chart depicts data from 2018 – 2020 from where it can interpret that DC had most fours which is 658 and KXIP had least fours which is 545. On looking into the sixes of league both KKR and MI have same number of sixes which is 359 and RR and SRH have least sixes which is again same 248.

**4. Highest scores by players**

The 10 highest scorer of IPL league are represented in this pivot chart from 2018 - 2020. KL Rahul scored highest runs in IPL (1922). At second place Shikar Dhawan have 1636. The pattern in chart shows the difference between first two highest scorer of runs is more than 200 but after that the difference between players scores is less. The least runs are made by David Warner (1240).

Figure 13. Highest Run Scorers

**6. Total boundaries vs total runs**

Figure 14. Total Boundaries VS Total Runs

This chart depicts data from 2018 – 2020 from where it’s clear that MI have both highest runs (7451) and boundaries (1015). Both RCB and RR have same boundaries, but the runs scored by RR are more than RCB. Again, on detail analysis it can note that KKR have more boundaries 1004 than CSK 913 but have less runs than CSK.

**7. Batting Average of Highest Run Scorers**

Below graph shows about the batting average of highest run scorer.

Figure 15. Batting Average of Highest Run Scorers

Virat Kohli have highest run (5426), but his batting average is less than most of players. AB de Villers have highest batting average, but his runs are at second least (4414) for top 10 players.

**Bowling**

**1. Top Bowlers of the season**

Figure 16. TOP BOWLERS OF THE SEASON

The above chart shows the Top 10 bowlers of IPL (Indian Premier League) from 2018-2020. Y-axis of the chart shows name of the bowlers and on X-axis shows the number of times the bowlers have dismissed the batsman during the league. JJ Bumrah and Rashid Khan are leading with 63 and 58 dismissals.

**2. Top Bowlers of the season**

Figure 17. Sucess Rate of Bowling types in IPL

The above chart shows the success rate of any bowler in IPL which is measured by 2 factors Number of times bowlers have taken wickets (dismissed batsman) and economy rate (Number of runs given by bowlers per over, the lesser the value is the better). The chart depicts on the left Y axis shows the number of dismissals, right y axis shows the economy rate and X axis shows the type of bowlers in IPL. According to the chart left\_arm\_fast medium and Leg break googly are the most economic bowler types and right\_arm\_fast\_medium and right\_arm\_medium is highest wicket taking bowler types, but Leg break googly is overall best with 3rd highest wicket taking type with 2nd best economic rate.

**2. Percentage of wickets taken vs Overs**

Figure 18. %OF WICKETS TAKEN VS OVERS

The final chart takes input of the bowler’s name, and it shows the blower’s strength overs and in which overs a particular bowler has dismissed most of the batsman and can also see bowlers’ strength overs. The following chart shows the bowling record of top wickets taker of IPL JJ BUMRAH from 2018-2020 and according to the chart he has taken most of his wickets in 17th and 19th over of the games with 20% of his total wickets followed by 15th over with 9% dismisses which shows he is really good and effective bowler in the final phases of match where batsman look forward to score more runs.

**6. DATA CLEANING AND TRANSFORMATIONS**

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabelled. If data is incorrect, outcomes and algorithms are unreliable, even though they may look correct.

In this project many null values, duplicates and NA values identified, and it is processed as expected.

The following packages are used in Jupyter python for the data cleaning and data transformation

**Graphical user interface, text, application

Description automatically generated**

**1. Matches**

This matches table had quite number of NA values, following snippet shows that method column has the greatest number of NA/Null values. Most of NA are valid one which are interrelated with other columns.

**Table

Description automatically generated**

As per the below snippet, duplicates are identified in 4 columns, “Rising pune supergaints” is duplicated with spelling error. It is proceeded and corrected.

**A picture containing text

Description automatically generated**

The duplicated are removed in the 4 columns and are updated with relevant data and it is shown below

**Table

Description automatically generated with medium confidence**

Total number matches played by each team are not calculated properly which is segregated in 2 different columns as team1 and team2. New variable is created as total matches to store the total counts of each team.

**Text

Description automatically generated** Text

Description automatically generated with medium confidence

**Text

Description automatically generated**

The above snippet the updated information on total number of matches played by each team.

The updated information of Matches table and home and away win table is shown below, where null values,

duplicates are processed.

**Graphical user interface

Description automatically generated with low confidenceText

Description automatically generated**

**2. Bowling**

The data cleaning has been done in the python using jupyter notebook, the necessary libraries should be imported to proceed with the cleaning which includes pandas, NumPy, seaborn, matplotlib and so on, the csv file was imported and then the number of NA values was calculated at each column and also the dataset has been checked for null values which are present in different forms which can be null, Nan, N/A and others, then if the whole row was null such rows will be dropped but there’s no such data available in all\_season\_bowling\_card, also there was one null value present in the economyRate column of the data which was handled by interpolation.

**Table

Description automatically generated**

**Table

Description automatically generated**

When it comes to data transformation the href column holds no valuable data which can be used in further analysis hence that column has been removed from all\_season\_bowling\_card data and there are no further transformations done in the dataset.

**3.Batting**

Batting data is imported from csv files to pythonand details are printed using print to know about dataframe.

**Table

Description automatically generated**

Using the info function got all the detailed information regarding the data frame. Batting data includes 24 columns containing Boolean, float, integers and objects.

**Table

Description automatically generated**

The count, mean and other central tendencies are calculated using the describe function to know about insights of data.

**A picture containing graphical user interface

Description automatically generated**

To clean data firstly null values are checked using is null function and can-do sum, count of null values using sum or count function after isnull function. In this sum function is used to know the exact numbers of null values in data frame.

Table

Description automatically generated

After checking the null values it’s important to clean data for future use or visualization of data using various tools. In this for cleaning data link, commentary, shortText, runningover, runningscore, isNotOut, captain, minutes, match name and full name. To make data look clean some columns are renamed using rename function shown in below snipt.

Graphical user interface, text, application

Description automatically generated

**4. Players**

As per below snippet players data is loaded for the analysis and cleaning.

**A picture containing diagram

Description automatically generated**

In the snip below null values for the players data frame are checked and summed using sum function. These values are handled in next steps by removing DOB column. The null values for country and bowling skill are kept as it is because we need this column in our analysis. After cleaning data information is previewed using info () function.

Table

Description automatically generated with medium confidence

Graphical user interface, text, application, email

Description automatically generated

The snippet below shows the unique values for batting hand and bowling skill. The second snippet shows the count of unique values in the players data. The data for players is not much so not much cleaning is needed and can be used.

Text

Description automatically generated

Text

Description automatically generated

**6. Bowling - IPL 2008 – 2020 Ball by Ball, All Season Bowling Card**

Excel Sheet Used: IPL 2008 – 2020 Ball by Ball, All Season Bowling Card

* Data Transformation: The Only Deleted column is “href” as the column from IPL 2008-2020 Ball by Ball as data can’t be used from that column.
* There are NA values are conditional present in the table, but NA here doesn’t mean data is not available but shows the condition didn’t happen.
* In IPL 2008 to 2020 Ball by Ball column “Overs” there are lot of 0 values which will be considered as 20th Over of the game which is final set of throws in each innings.

**7.DATA ANALYSIS**

**1. BATTING DATA ANALYSIS**

**PROBLEM STATEMENT & OBJECTIVES**

* TO FIND TOP BATSMEN IN THE LEAGUE
* TO ANALYZE THE FACTORS BEHIND BATSMAN’S PERFORMANCE
* TO EVALUATE PLAYER’S PERFORMANCE IN DIFFERENT CONDITIONS

In order to find the solution for the problem statements above data regarding batting will be explored and analysed with the help of visualisation using Tableau. The following graphs will answer those problems step by step and will provide insight of performances of batsmen.

**1. Players with Most Runs and Matches**

In T20 Cricket, fans love high scoring games which are more interesting and enjoyable. These high scoring matches depend on top quality performances of batsmen. Let’s evaluate who has scored the maximum runs in IPL since it commences.

Chart, bar chart

Description automatically generated

Figure 19. Top 10 All-Time Run Scorers along with the Total Number of matches they have played

From the above plot, it is clear that Virat Kohli is the highest run scorer with 5,878 runs in 184 matches. While David Warner has played least number of matches (142), he is still 3rd highest run scorer. It can be deduced that David Warner records high runs in each game in comparison to others and is a highly bankable player to expect high scoring games when he’s playing.

**2. Analysing the latest Performance of Batsmen**

Some of these players have retired or will be retiring in next few years. Keeping the future prospects in mind, data is analysed from recent years. Below graphs are based on data from IPL Season 2017-2020.

Chart, bar chart

Description automatically generated

Figure 20.Top Run Scorers and their runs

In the above graph, top 10 batsmen are arranged based on total runs scored between year 2017-2020. Only 4 Batsmen from *Figure 1* made it in this graph. This shows that either performance of rest of the players has significantly dropped or they have retired. Interestingly, David Warner is still at 3rd position with 1881 runs, while Shikhar Dhawan has taken 1st rank with 2115 runs.

Let’s look at the impact of boundaries for these top scorers.

Chart, timeline, bar chart

Description automatically generated

Figure 21. Number of Boundaries. Total number of ‘Fours’ are mentioned at the end of bar while Total number of ‘Sixes’ are depicted by a black vertical line on each bar and colour range.

From the graph above, it is evident that Shikhar Dhawan mostly goes for a ’4’ than a ‘6’. With 243 ‘fours’, he is way ahead of his competitors in number of boundaries, even though he has hit the second least number of ‘sixes’ i.e., 46. The most balanced number of boundaries is of KL Rahul with173 ‘fours’ and 80 ‘sixes’, making him the 2nd highest run scorer (1,922 runs).

**3. Top 10 Players**

The chart below shows the balls faced by top 10 players in different stadium. The second chart shoes the balls faced by players only without venue.Graphical user interface, table

Description automatically generated

Figure 22. Top 10 Player based on Stadium Chart, timeline, bar chart

Description automatically generated

Figure 24 . Count of ball Faced

**4. Analysing Performance of Top Players in Different Venues**

IPL matches are played at multiple locations, In India and out of India as well, such as U.A.E and South Africa. Let’s analyse how top batsman have fared in these venues since 2017.

Chart, bar chart

Description automatically generated

Figure 25. Top 5 Batsmen's score categorised according to various venues

From the graph, it can be seen that every batsman has scored high runs in any one particular stadium e.g., DA Warner has scored 769 runs in Hyderabad RR Pant has scored 743 runs in Delhi. That is because these grounds are their home grounds. Although, S Dhawan and KL Rahul have performed well at many other venues as well, which signifies that they are more bankable players.

**5. Dismissal Analysis**

**Chart

Description automatically generated**

Figure 26. Dismissal of Feilders

The above chart depicts the top 10 fielder stats based on the different kinds of dismissal kind which are bowled, caught, caught and bowled, run out, stumped and so on and the NA values are excluded from the data because not all balls contributed to a wicket by a particular fielder.

**Data Analysis on Bowling**

The below data analysis shows the number of analyses done on the bowlers and type of bowlers. With the help of this analysis, we are trying to find out who succeeds and what type of bowler would likely to perform.

The Following Results have been Concluded from Bowling analysis:

* Having a lower Economic Rate or Higher number of wickets is good shows the competency of the bowler but best bowlers are the those who have lower economy rates with high amount of wicket for Eg: Piyush Chawla has 2nd most wicket taking bowler also high in conceding runs which makes him decent bowler but R. Ashwin has highest runs conceded but not in Top 5 list of wicket taking bowlers shows poor form.
* In Wickets and Runs conceded not only affects the bowlers but also to the team it can be seen in graphs that whenever the number of runs go down the number of wickets is high. In 17th,18th,19th which also shows in the final phases of the game the batting team plays more aggressively and careless about wickets and batting teams are more cautious in 5th and 6th overs.

**1. Bowlers Conceded Most Runs and Top wicket Takers:**

Graphical user interface, text, application

Description automatically generatedGraphical user interface, application

Description automatically generated

Above is the list of bowlers who have given the greatest number of boundaries in a given about of seasons in which Bhuvneshwar Kumar is leading with 92 and Umesh Yadav with 90.

Above is the list of bowlers who have taken the greatest number of wickets in a given about of seasons, higher the number of wickets better the bowler, in which Dwayne Bravo is leading with 31 and Piyush Chawla with 30.

**2. List of Bowlers Conceding Most Runs and Top Bowlers Conceded Most number of 6’s:**

Graphical user interface, application

Description automatically generated Graphical user interface

Description automatically generated

This chart shows the number of bowlers who have conceded the greatest number of runs. This value affects the performance of the bowler negatively, as number of runs given by bowlers should always be less.

Number of 6’s is the bigger than 4’s this list shows the bowlers who have given the most sixes (Scoring six runs in single Ball) which can also affect the performance of bowler in a bad way. Piyush Chawla leads the here with 46 sixes and Jadeja with 39 on 2nd spot.

**3. Economy Rate by Playground:**

Map

Description automatically generated

The above is the geo map which shows the city (in which matches have been played) by the economy rates of bowlers. This visualization shows amount of runs bowlers tend the concede in given number of overs in the marked playground. The more economy value the higher chance of bowlers giving high number of runs.

**4. Fall of wickets By Over:**

Chart, bar chart

Description automatically generated

The above charts shows the amount of wickets going down by overs in which shows in 19th over of the game we can expect highest amount of batsman getting out and lowest chances of getting wickets for bowlers is 9th over with only 62 wickets.

**5. Most Runs Given by each over:**

Chart, line chart

Description automatically generated

The chart is similar to the chart we saw above showing the amount of runs conceded by the over, also shows where most likely bowlers tend to give more runs.

**7. Economy Rate by bowling styles:**

Table

Description automatically generated

IPL hosts many types of bowlers, the list above shows the types of bowlers arranged by lowest to highest, which shows the amount of runs particular type of bowler gives while bowling.

**TABLEAU DASHBOARD**

The Tableau dashboard is created with 5 different dashboard which is merged together as website, can be navigated though a Navigation tabs.

This Dashboard contains a full detail of the IPL data Analysis. It is published in Tableau online, refer the below link

[**https://public.tableau.com/app/profile/vignesh.dharuman/viz/IPLDATA\_16496053335650/MainDashboard?publish=yes**](https://public.tableau.com/app/profile/vignesh.dharuman/viz/IPLDATA_16496053335650/MainDashboard?publish=yes)

**8. EXPECTED RESULTS**

The project will yield three major deliverables:

**Methods** **developed:**

In this team will provide detailed summary of the methods used, how they are chosen to do a specific task and how they are used to get best result out of the dataset. Team will also provide insight of how methods are used to know about relation between csv sheets, tableau, and visualization with help of data analysis.

**Model to predict:**

Project will deliver the various ways to correlate variables of data and will provide a great overview of data. Our team will also provide different charts in form of visualization to understand analysis easily for future reference.

**Analysis of data**:

In the end this project will provide a well-managed dashboard on tableau with reference to other documents and material used to carry out this project. our team will provide a detailed narrative summary of each step of project.

**9. References**

**Datasets links**

[**https://www.kaggle.com/ramjidoolla/ipl-data-set**](https://www.kaggle.com/ramjidoolla/ipl-data-set)

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

[**https://www.kaggle.com/datasets/rajsengo/indian-premier-league-ipl-all-seasons?select=all\_season\_batting\_card.csv**](https://www.kaggle.com/datasets/rajsengo/indian-premier-league-ipl-all-seasons?select=all_season_batting_card.csv)

[**https://www.kaggle.com/datasets/rajsengo/indian-premier-league-ipl-all-seasons?select=all\_season\_bowling\_card.csv**](https://www.kaggle.com/datasets/rajsengo/indian-premier-league-ipl-all-seasons?select=all_season_bowling_card.csv)

**Document preparation references**

https://www.pngkit.com/view/u2e6y3t4e6u2q8u2\_data-analytics-methodology-diseo-de-una-infografia

https://www.researchgate.net/figure/The-fundamental-steps-of-the-exploratory-data-analysis-process\_fig3\_32993077