1.Introduction

Motivation

The key aspect of any sport or a game, in general, is to keep track of certain metrics, which eventually decide the fate of a team or a player.

For example, event management decides to appoint a person named X to keep track of the metrics related to gameplay, time, score during the whole length of a basketball match happening back in times where there is no technology to digitalize the tracking of score and time, and the whole event management including players rely on X to decide the winner.
In a case where X unsees a pointer or gets distracted or lets clock overrun than stipulated time, there might be a change in the winning fate of a team, especially when the match is neck to neck.

In any sport, tracking of gameplay metrics at every second is crucial. Not only at a team level, but at a player level as well to keep track of records and achievements. With the metrics tracked and data generated during the whole time of the gameplay, most of the sports teams, and their sponsors are trying to get insights out of the sports data that help plan for the upcoming matches/series and get a rough understanding of what to expect. (Schroer, Alyssa)

- Why cricket ?

    - Cricket is one of the unique sports with a play style that is segmented as plays are usually on a ball-by-ball basis, this provides an opportunity to dynamically strategize in-game. Games such as soccer & Basketball which are more continuous in nature cannot take advantage of such analysis or strategies

    - The visualizations shown below (Fig 1.2a & Fig 1.2b) are some of the common visualizations shown while watching the game, however they are at the specific match level and do not show us details at a larger scope.

    - We believe that with the extremely high volume of cricket data being generated and collected nowadays (530k+ players, 540k+ matches annually) there is a massive scope for visual analytics at a larger scale.

    - Cricket is one of largest and most popular sports in India, as all the team members of our group belong to the Indian subcontinent we were all avid cricket fans from our childhood and this further convinced us to choose the sport of cricket for our visualization project.

Fig.1.1a: Introduction to the game of cricket

Fig.1.1 b: Major formats in the game of cricket

Fig.1a explains the game of cricket describing how the scoring, dismissals and the flow of the game looks like

Fig.1b explains the major formats involved in the game of cricket played internationally

Here are basic terms used in cricket -

- Batting: The act of hitting the ball with the bat.
- Bowling: The act of throwing the ball to the batsman to get them out.
- Batsman: A player who is currently batting.
- Bowler: A player who is currently bowling.
- Wicket: A set of three wooden stumps (and two bails on top) that the bowler tries to hit with the ball to get the batsman out.
- Pitch: The rectangular area of the cricket field where the game is played.
- Innings: A single team's turn to bat and score runs.
- Over: Six consecutive balls delivered by the same bowler.
- Run: The number of bases a batsman advance by hitting the ball.
- Boundary: The edge of the playing area. If the ball is hit and crosses the boundary, the batting team is awarded a certain number of runs.
- Duck: When a batsman is out without scoring any runs.
- Hat-trick: When a bowler takes three wickets in three consecutive deliveries.
- No ball: An illegal delivery by the bowler, which results in one extra run for the batting team and another delivery.
- Wide: A delivery that is too far outside the batsman's reach, which results in one extra run for the batting team and another delivery.

Cricket is one of the most liked, played, encouraged, and exciting sports in today's time.
It's the second most popular game in the world with 2.5 billion fans worldwide(Veroutsos, Eleni). With the increasing number of matches with time, the data related to cricket matches and individual players are increasing rapidly. Data in Cricket is generated every day for 365 days. With the ball-by-ball information of 530k+ cricket players in close to 540k+ cricket matches at approximately 12000 cricket grounds across the world - this data is extremely voluminous (*ProjectPro).*

Moreover, the need to use data analytics and the opportunities to utilize this data effectively in many beneficial ways are also growing, such as the selection process of players in the team, predicting the winner of the match, circumstances that led to a particular outcome, and many more.

Existing work

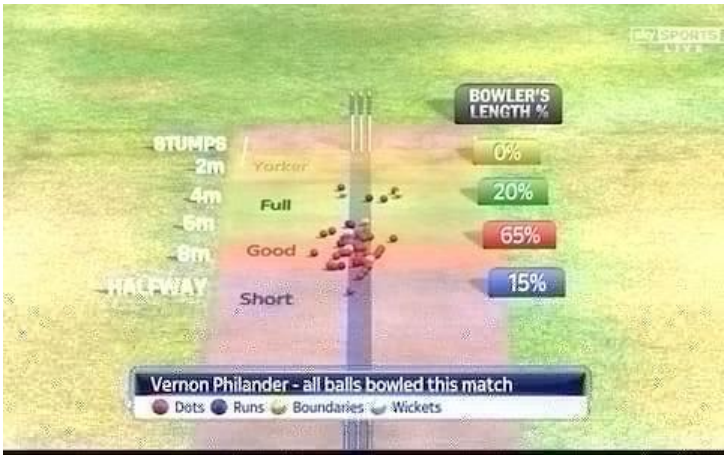


Fig.1.2a: Distribution of balls pitched split by bowler's length

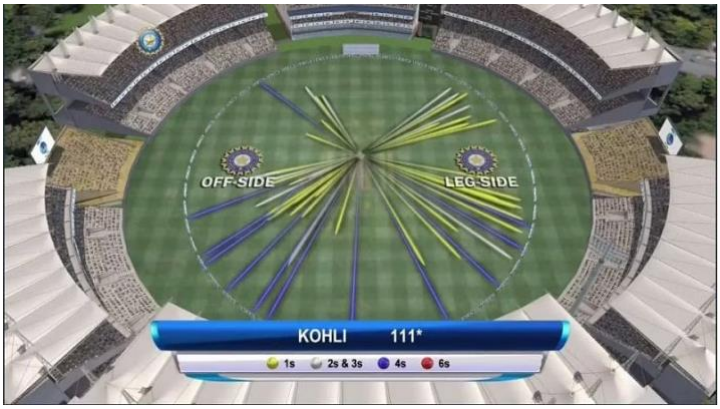


Fig.1.2b: Virat Kohli's Wagon Wheel

Fig 1.2a and Fig 1.2b are the most common visualizations shown during the live cricket match (Jain, Sahil).

Fig.1.2a shows the length ratio of the number of balls pitched from the bowling split across the length of the ball pitched from the bowling end.

Fig.1.2b is the wagon wheel visualizations where it showcases the shots hit by batsman towards the off-side and leg side.

Home factor usually plays a major role in deciding the winning fate of the match. Here are some existing visualizations that showcase the effect.

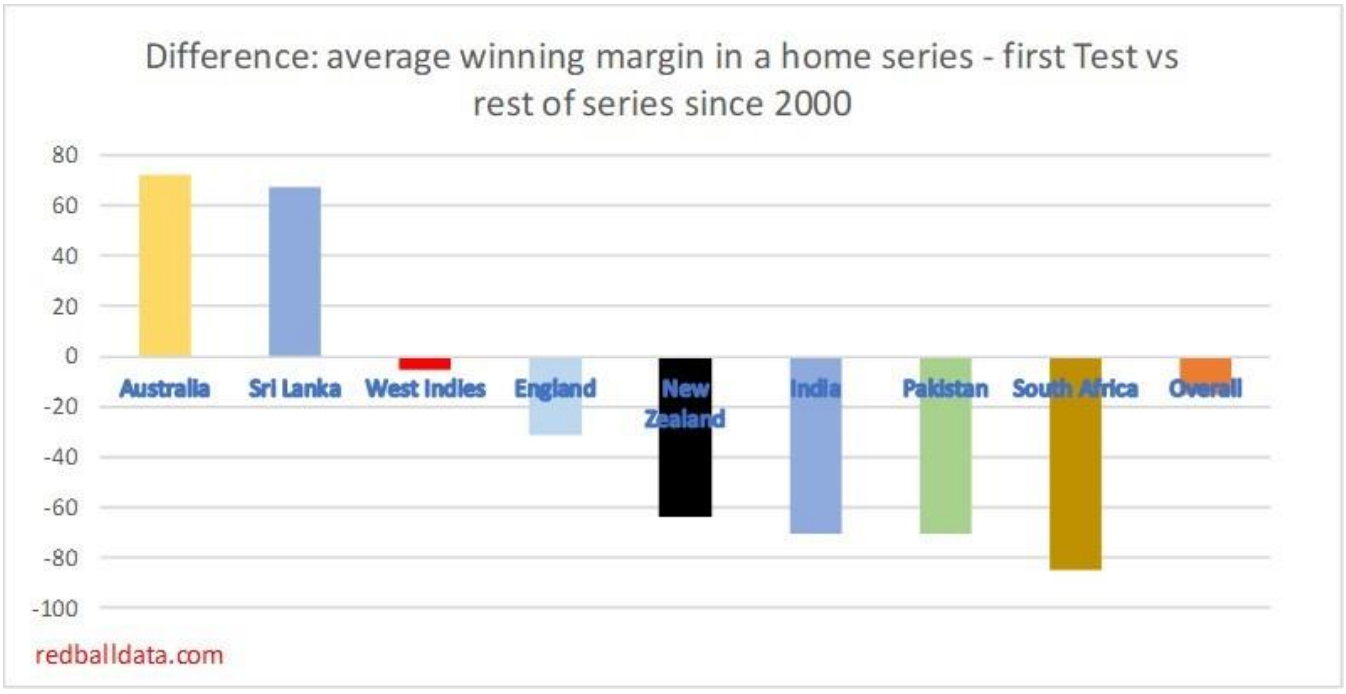


Fig.1.3: Relative home advantage

Fig.1.3 shows the relative home advantage in the first Test of a 3+ Test series as compared to the rest of that series

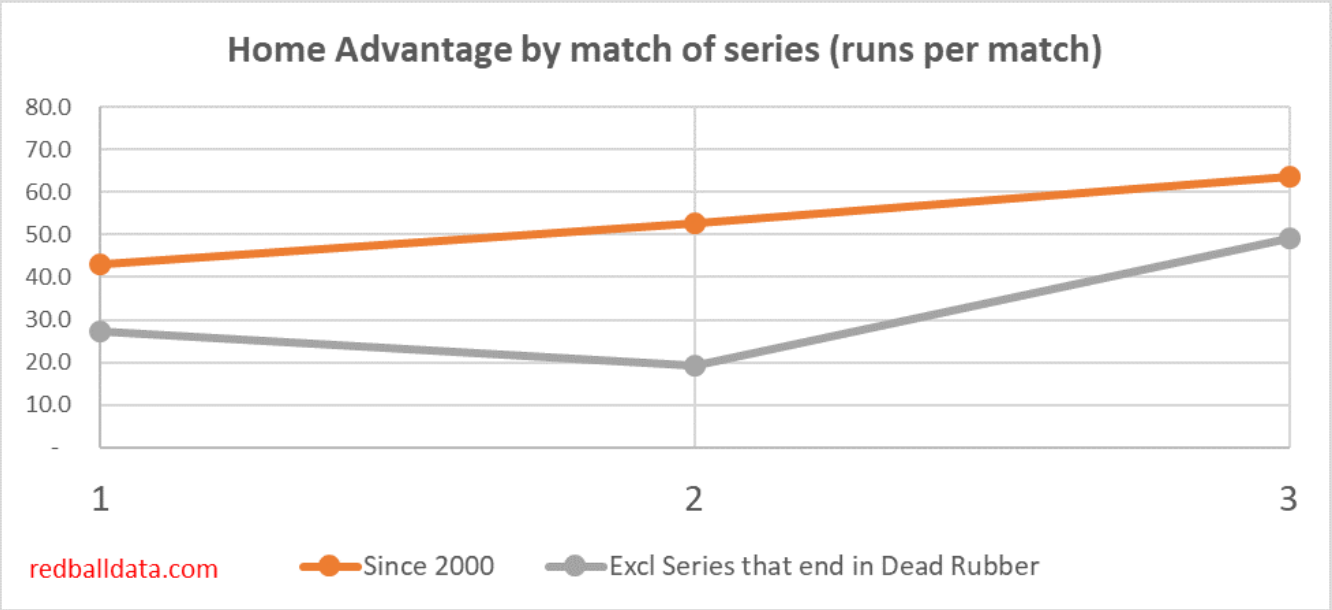


Fig.1.4: Home advantage measured in runs

Fig.1.4 shows the home advantage measured in runs, both including and excluding series that are decided before the last match of the series.

Excluding one-sided series shows lower home advantage (because it excludes big home wins when a visiting team can't compete with a superior host team). The overall effect is the same though- home advantage gets markedly bigger in the later Tests (*Red Ball Data).*

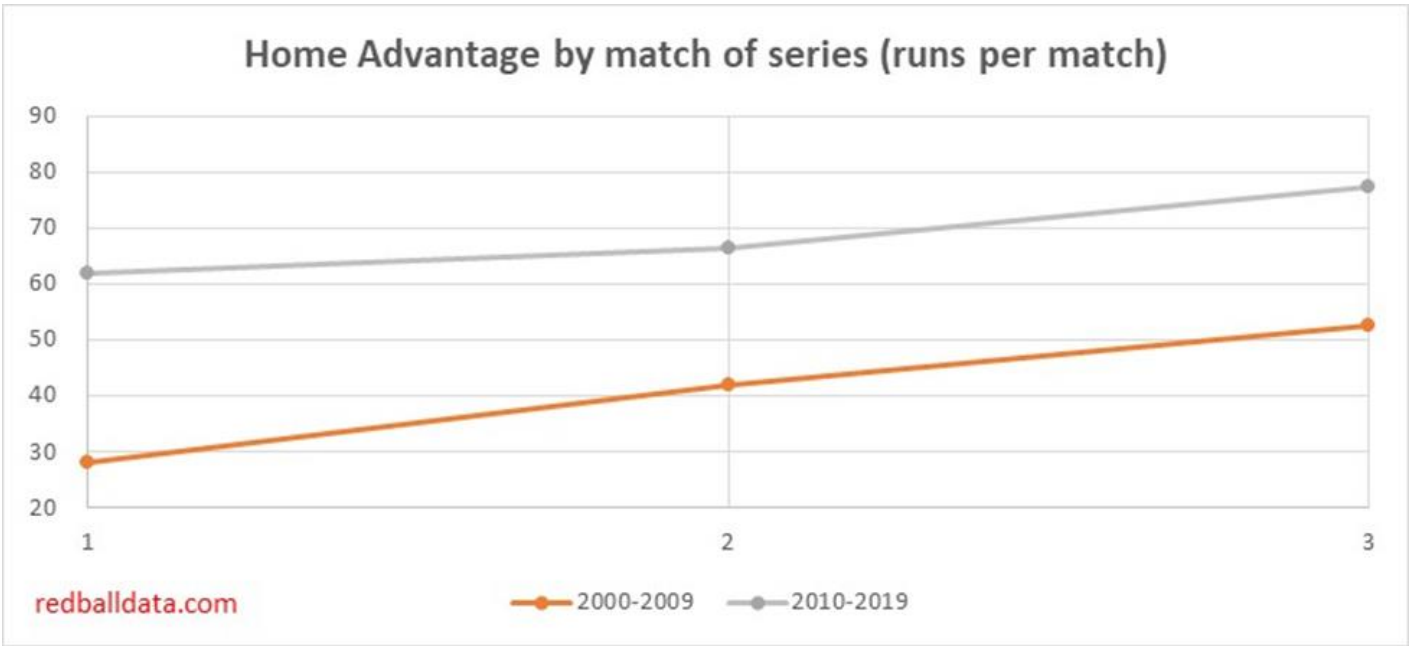


Fig.1.5: Home advantage in runs by match of series

Fig.1.5 shows the home advantage by match of series based on the runs margin. A series consists of 3+ matches.

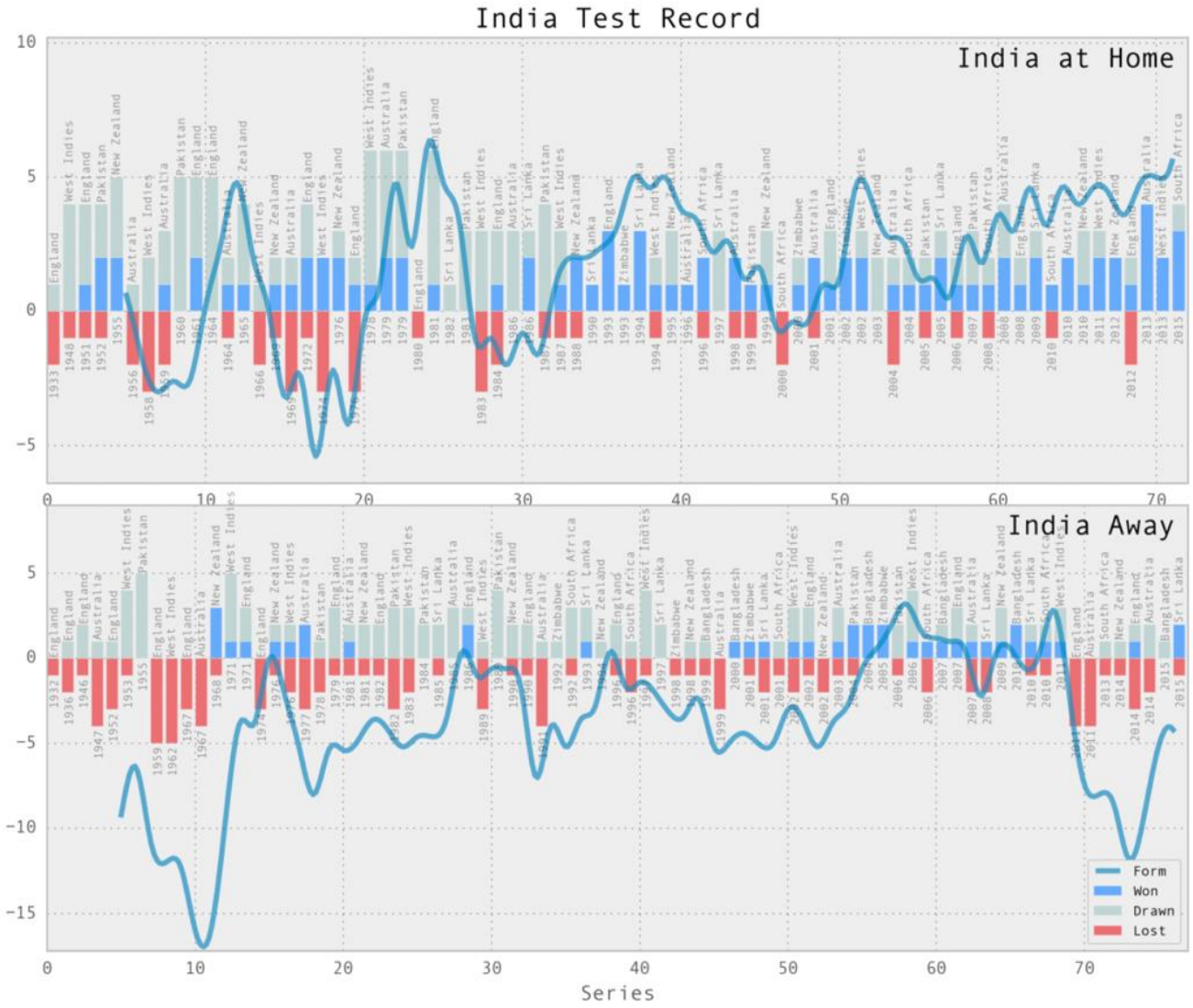


Fig.1.6: Home pitch factor influence on team India in international test cricket

Fig.1.6 is an interesting visualization of how the home pitch factor influenced the form and its winning for team India in international test cricket.

While India is playing in one of its home pitches, its form is positive and the wins, ties were higher when compared to losses.

While away from India, the losses were higher compared to wins and consecutively the form also was not very good. These trends can be observed clearly in the time ranges of 2000 – 2003 and 2008-2015.

Indian Premier League(IPL) matches are the most viewed league matches in the game of cricket around the world. (Kreedon) IPL has become so popular that the TV viewership for the 2021 season crossed 400 million which is almost half of the total number of TV viewers in India (around 836 million). The popularity of IPL keeps on rising as it is not only being followed in India but also in various other countries such as Sri Lanka, Bangladesh, UAE, etc.

Here are some of the existing visualizations that show the correlation for Mumbai Indians between most matches won vs toss winning frequency:

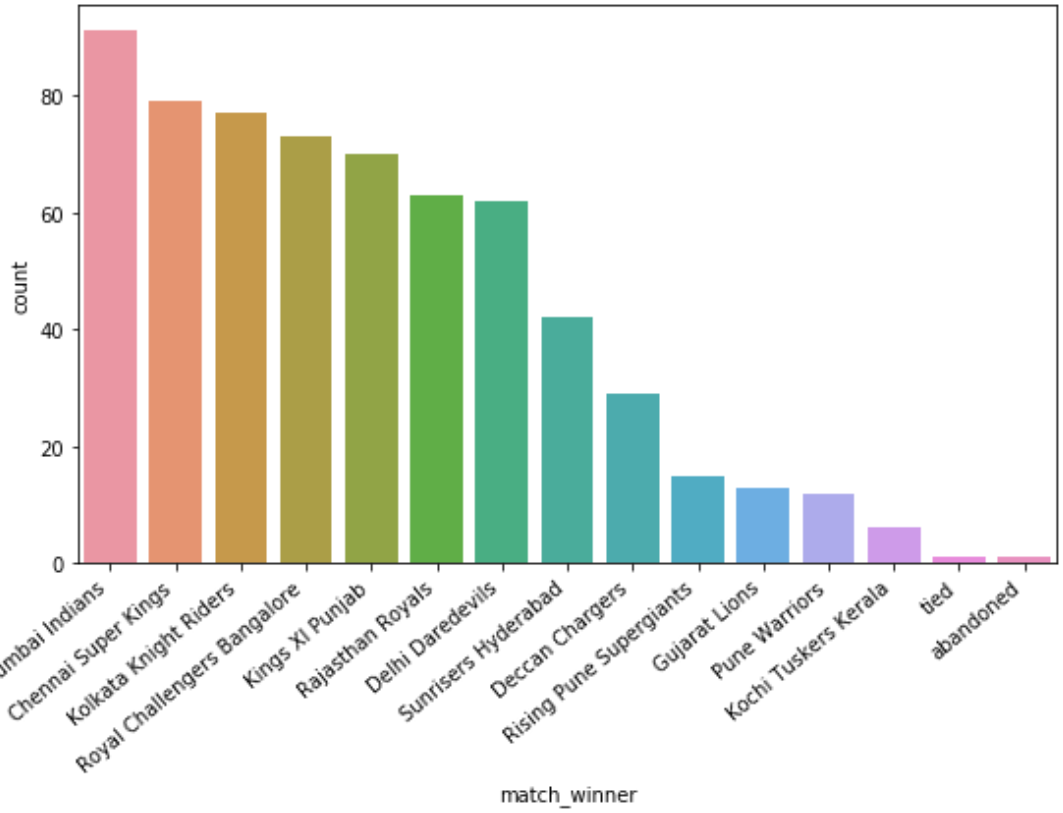


Fig.1.6a: Match winning count frequency

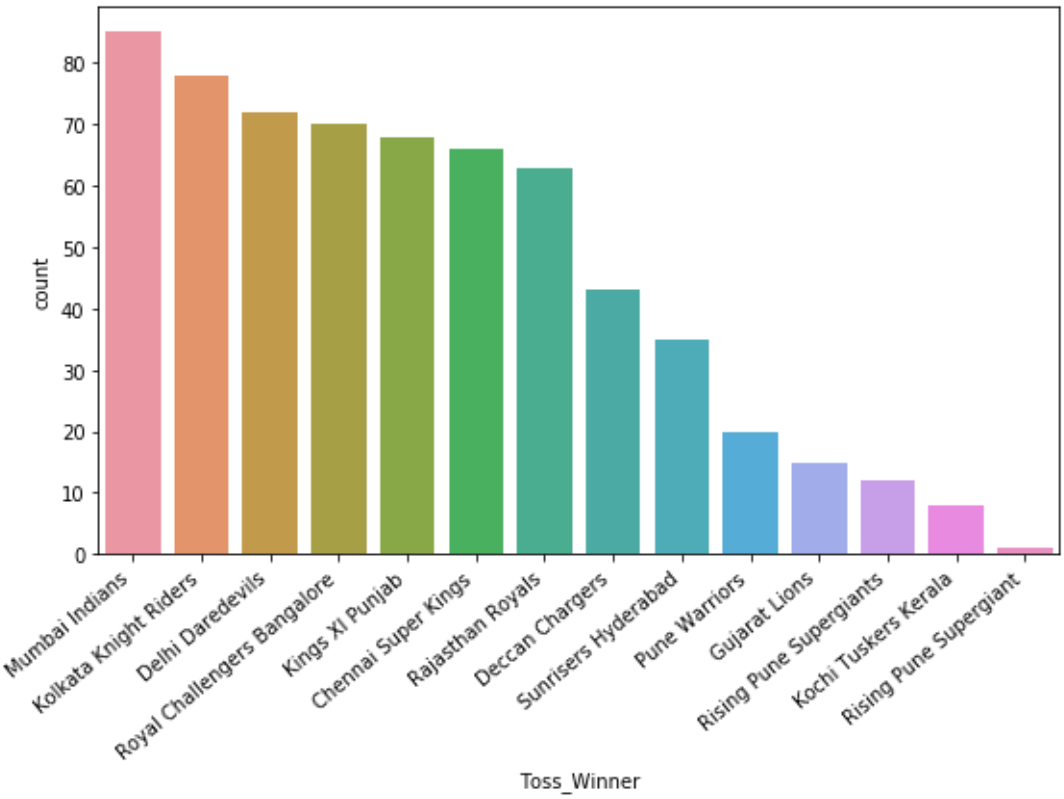


Fig.1.6b: Toss winning count frequency

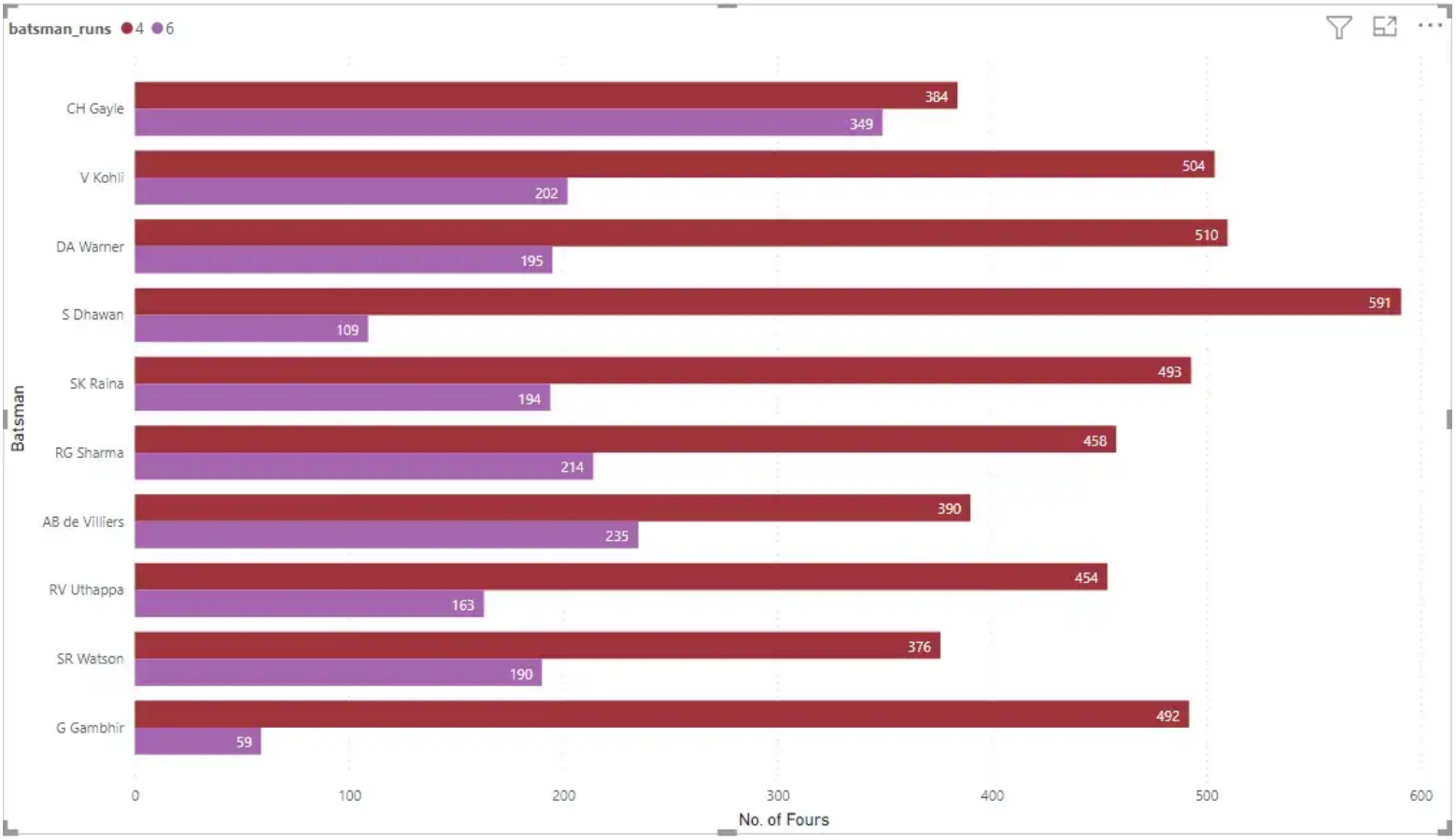


Fig.1.7: Top 10 players VS boundaries

Fig.1.7 shows the boundaries for top 10 players in IPL matches, which are ideally the most enjoyable and lucrative moments for fans and the team respectively (Singhal, Shashank).

Here the bars in pink, represent the number of 6s and bars in red represent the number of 4s by the respective batsman

There are some players who top the game of cricket consistently. It is interesting to see how this consistency is being displayed.

Here are some interesting visualizations about top performing players Messi, Ronaldo in soccer.

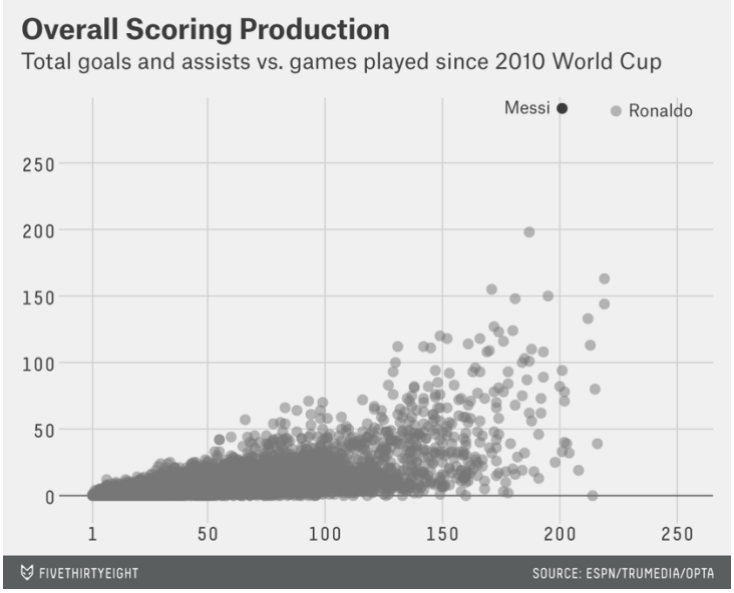


Fig.1.8: Overall Scoring Production

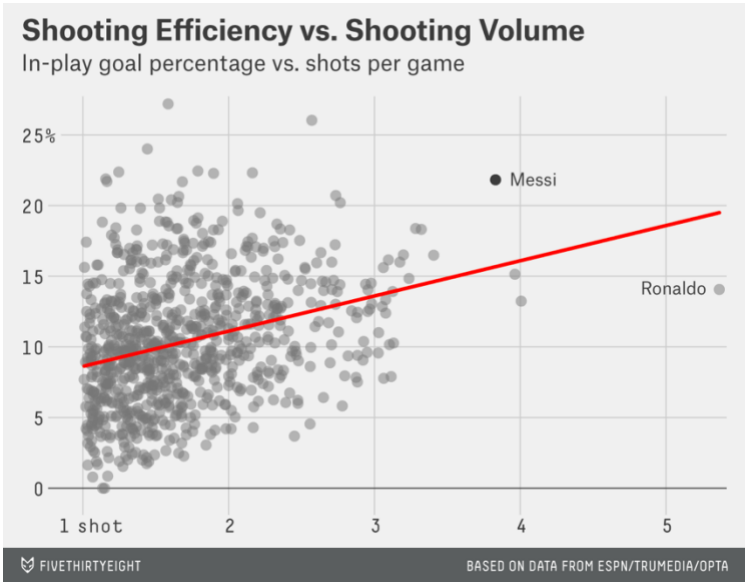


Fig.9: Shooting efficiency vs shooting volume

Fig.1.8 shows overall goals and assists across the number of games played since 2010 World cup.
Goals and assists are important because they directly contribute to a team's success. In most football leagues, teams are awarded three points for a win, one point for a draw, and no points for a loss. Scoring goals is the most effective way for a team to win a match, so players who can score goals are highly valued.

Assists are also important because they can lead to goals. A player who can create scoring opportunities for their teammates can help the team to score more goals, which can increase the team's chances of winning. Some of the greatest football players in history, such as Lionel Messi and Cristiano Ronaldo, are known for their ability to score goals and create assists.

Here the scatter plots are used to define the density of the goals and assists for Messi and Ronaldo represented in dark shade and lighter shade respectively.

Fig.1.9 shows the shooting efficiency vs shooting volume for Messi vs Ronaldo in soccer game on a scatter plot to visualize the comparison patterns of shooting for these 2 respective players.

We can infer that Ronaldo has a visibly small percentage of shooting efficiency and shooting volume higher than Messi.

Contribution

Given the scale of cricket game's data and considering the visualizations mentioned in the existing work section, we have come up with the intent to understand three different datasets that would initially help in conducting analysis on the following points respectively:

- Team's performance in home and far from home pitches
- Top players' data along with their records in batting
- IPL records

These datasets are chosen in order to analyze data at 3 different levels of hierarchy – overall team's performance in the international matches, league records of the teams involved to explore at ball-by-ball level and to understand the granular level of the data – player's statistics.

A new set of findings/insights we are planning to generate are based on the following key questions:

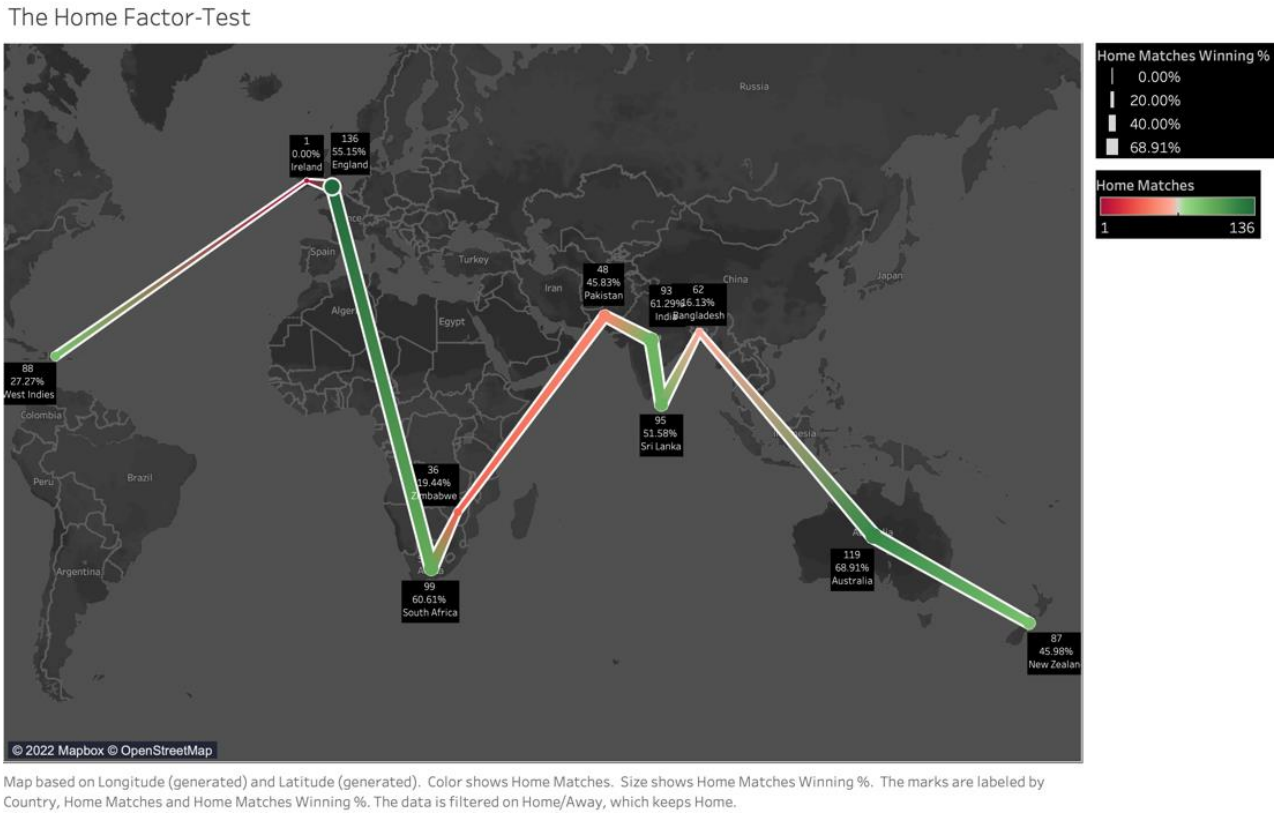
- How does the influence of home venue decide the team's winning fate at international level?
- How is home/away venue factor related to winning the game?
- How does the toss decide the winning of the game in IPL?
- How Sachin Tendulkar's and Virat Kohli's (two of the consistently top performers in cricket in different time periods who have a very thin time overlap while playing) gameplay is different based on:
    - Runs scored vs Balls Faced
    - Runs scored vs Minutes spent in the field
- How number of boundaries influence winning at team level in IPL?
- How does the distribution of runs determine the win or loss in IPL match?

2. Data & Methods:

Ideas, sketches, prototypes

Since we have listed out the key questions, here are the visualization ideas to achieve what we are trying to showcase based on the type of question.

- **Key Question**: How does the influence of home venue decide the team's winning fate at international level?
  in order to showcase the winning fate of the match, we have broken this problem into 3 parts as showcased in Ref. image 1:
- Initially showcase the countries that are actively participating in the game of cricket across the world to understand which parts of the world play cricket actively
  - World map is used to point the countries involved in the game
- Further, showcase the frequency of games hosted by each of these countries
  - Red-Green gradient on the world map at the respective country to represent the number of matches hosted



Ref. image 1

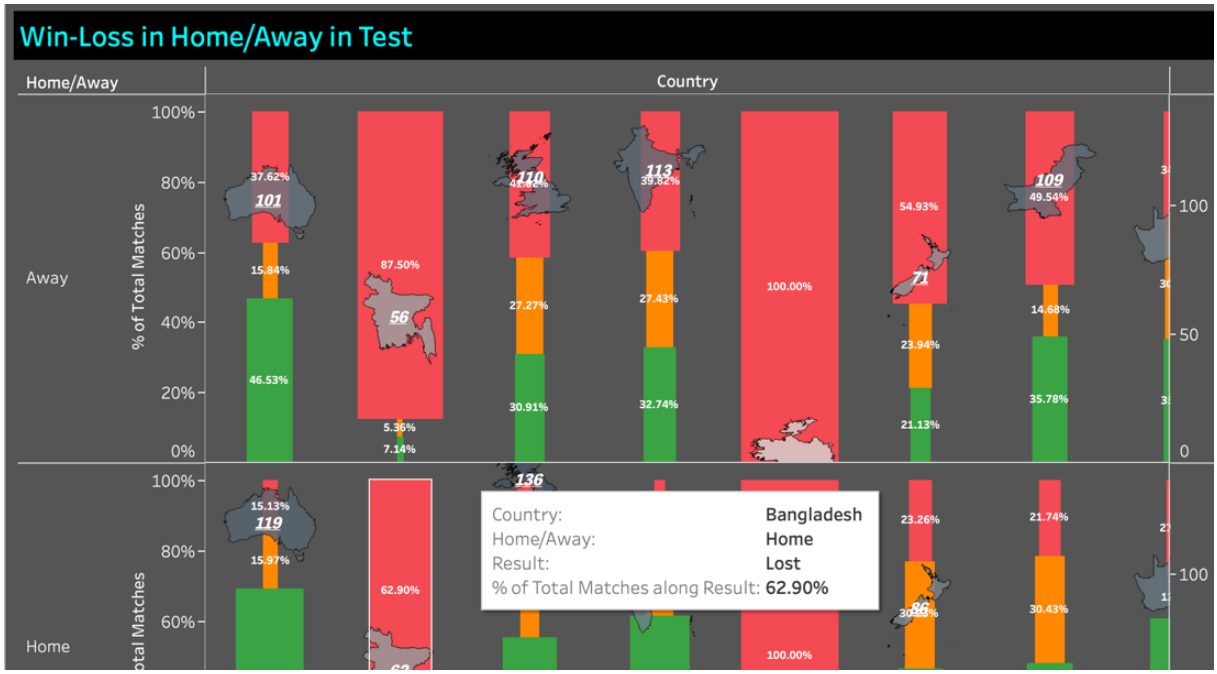
- Eventually, display the winning % of the matches hosted by the respective country to quantify the impact of the home factor on winning a match
  - Size of the point is used to represent the % of winning matches out of the matches hosted.

We have planned to create another set of visualizations we have used to represent the comparison of Home/Away factors in winning the match

- **Key Question:** How is home/away venue factor related to winning the game?

  In order to draw a comparison of the home/away factor we have decided to divide the Home/away factor into 2 different charts: the Home and Away categories and represent the winning/losing % of the matches played in red/green respectively.

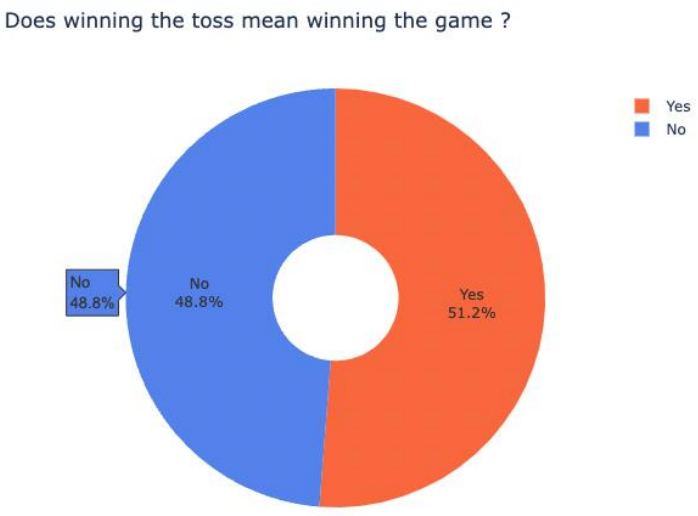
  Further winning/losing % of the matches played in the home/away venue is varied based on the size as well to quantify the proportion of winning rate. (Ref. image 2)



Ref. image 2

- **Key Question:** How does the toss decide the fate of the winning game in IPL?

  Initially to answer this question, we have planned to visualize the toss decision across the years and then finally understand winning fate of game abased on toss (Ref. Image 3)



Ref. image 3

❖ **Key Question:** Does Sachin Tendulkar and Virat Kohli (two of the consistently top performers in cricket in different time periods who have a very thin time overlap while playing) have a different gameplay?

In order to answer this question, we have taken several factors:

- Distribution of runs scored (Histogram of runs scored in match)

- Runs scored vs balls faced
    - Since this analysis is a bivariate type, we have chosen scatter plot for the runs scored by plotting the balls faced on the X-axis

- Runs scored vs minutes spent
    - Since this analysis is a bivariate type, we have chosen scatter plot for the runs scored by plotting the minutes spent on the X-axis

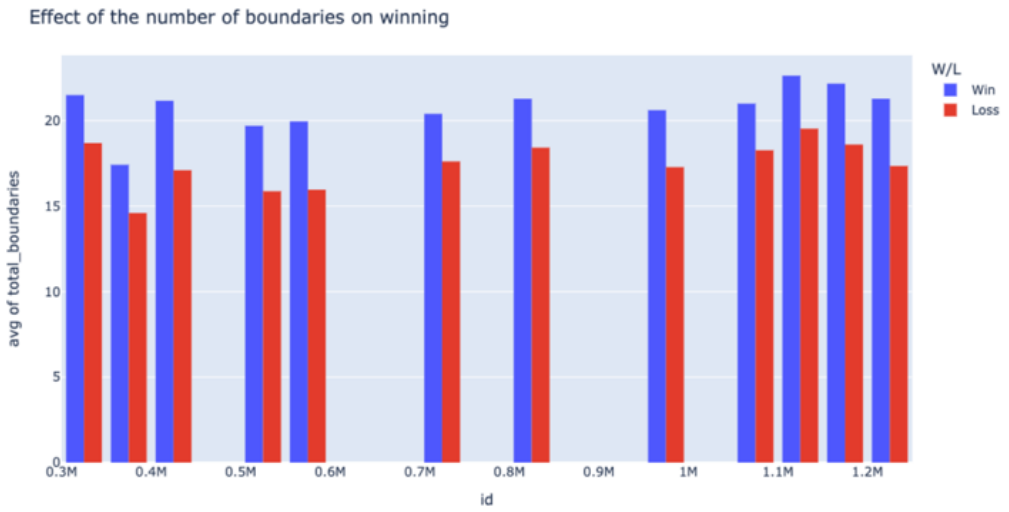
- Matches played year wise
    - Line chart is used here to understand the trend of the matches played yearwise

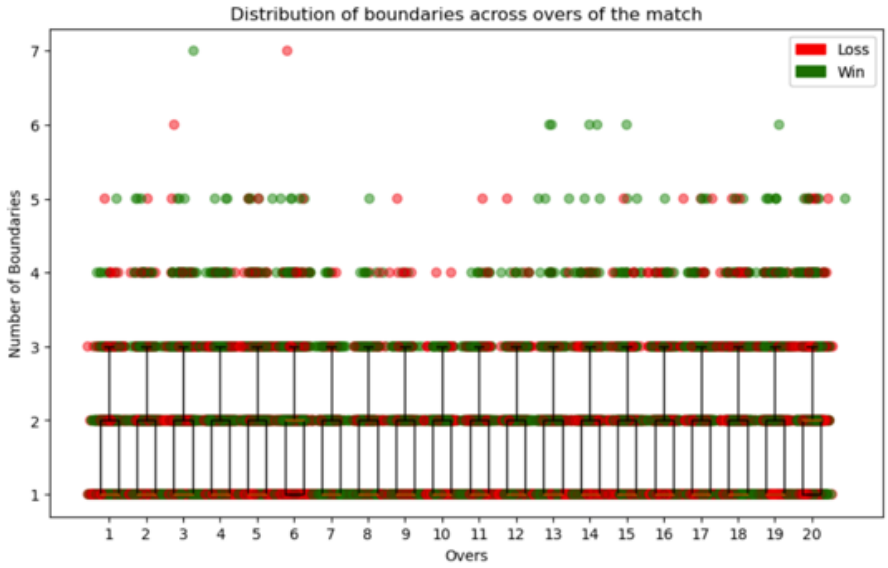
- Strike rate comparison:
    - In cricket, the strike rate is a measure of how often a batsman scores runs. It is calculated by dividing the number of runs scored by the number of balls faced, and it is typically expressed as runs per 100 balls. For example, if a batsman scored 50 runs in 25 balls, their strike rate would be 200 (50 runs / 25 balls * 100 = 200).
    - Histogram of strike rate comparison is done for Sachin vs Kohli for drawing comparison of the gameplay

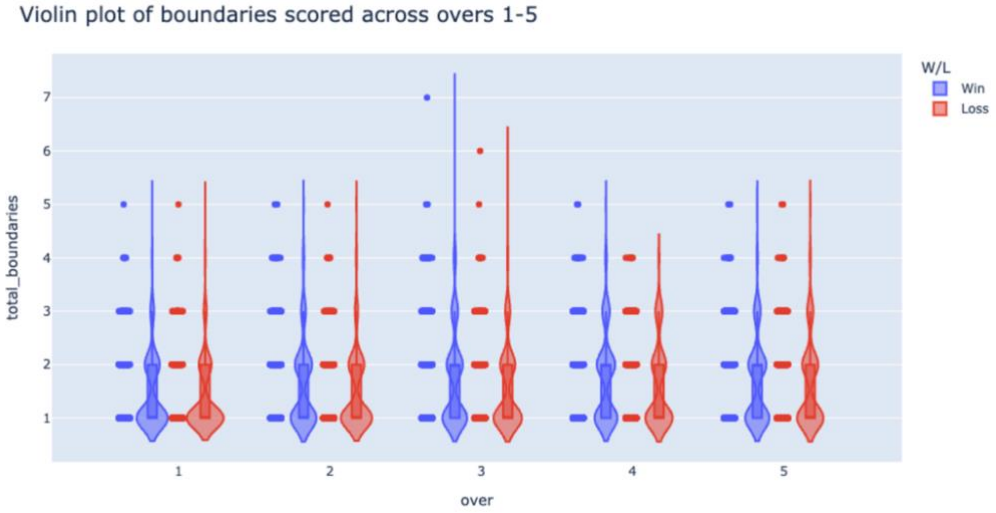
❖ **Key Question:** How number of boundaries influence winning at team level in IPL?

To understand this data, we have made use of 3 different types of visualizations:

- Effect of number of boundaries on winning (Ref. image 4)
    - Average boundaries across the win/loss
    - Distribution of boundaries across win/loss
    - Violin plot of boundaries scored across overs 1-5, 5-10,10-15,15-20 segments. Here the density/spread of the violin plot at the base shows the number of boundaries scored

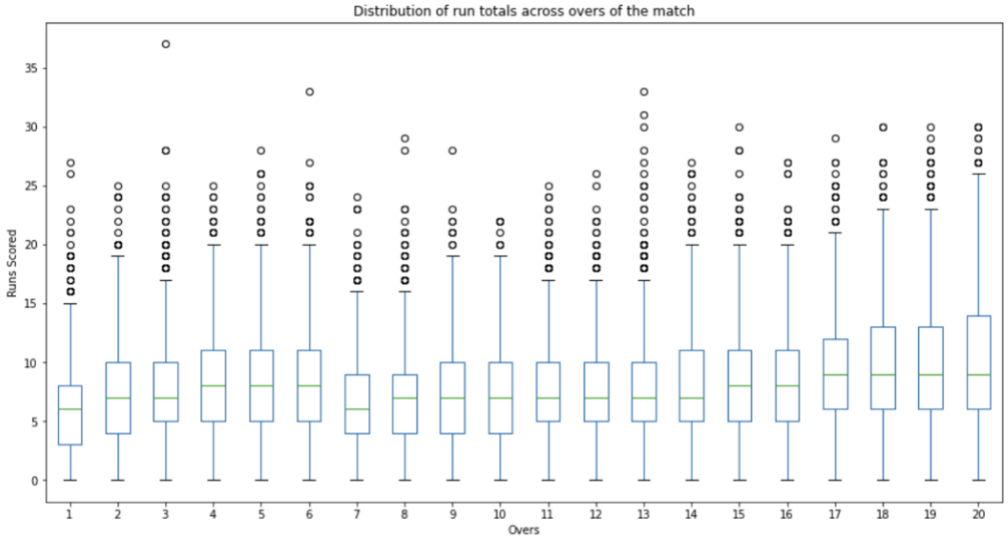






Ref. image 4

Key Question: How does the distribution of runs determine the win or loss in IPL match?
In order to answer this question, we have chosen the boxplot showing the runs scored across the nth over from (1 to 20) in a IPL match. (Ref. image 5)



Ref. image 5

Data & Methods: visualization methods selection

- **Scatterplots**

For visualizing ball-by-ball data in cricket, one option is to use a scatter plot. A scatter plot plots the runs scored on the y-axis and the ball number on the x-axis, but instead of connecting the data points with a line, it shows them as individual points. This can be useful for seeing the distribution of runs scored in each ball and for identifying any outliers (balls where a large number of runs were scored).

A scatter plot is a type of graph that is used to visualize the relationship between two numerical variables. It plots the values of one variable on the x-axis and the values of the other variable on the y-axis, and shows the data points as individual dots on the graph.

There are several advantages to using scatter plots. One advantage is that they can show the relationship between two variables in a straightforward and intuitive way. For example, if you have data on the height and weight of a group of people, a scatter plot can quickly show you if there is a positive, negative, or no relationship between the two variables.

Another advantage of scatter plots is that they can show the distribution of the data. For example, if you have data on the runs scored by different batsmen in a cricket team, a scatter plot can show you if the runs are distributed evenly across the batsmen or if some batsmen score more runs than others.
However, there are also some disadvantages to using scatter plots.

One disadvantage is that they can become cluttered and difficult to interpret if there are a large number of data points. This can make it difficult to see the overall pattern or relationship in the data. Additionally, scatter plots do not show any summary statistics or other information about the data, so you need to interpret the data visually.

- **Histogram**

Histograms can be used for charting continuous frequency distribution. These work well with large ranges of information. These histograms are especially useful when dealing with value ranges like runs scored by a player, Strike rate of players e.t.c.

A single glance at a histogram gives us some idea about the shape and spread of data. When compared to box plot one of the disadvantages of histogram is that it cannot give us extra information like median, upper quartile, and lower quartile of data. Histograms cannot be used when comparing multiple categories and they use a lot of ink and space to display very little information.

- **Line chart**

The line chart is used to display series of data points connected by straight solid line segments. For example, in this case we used line charts to analyze how the number of matches played by players varies with time (year by year during their career). These charts work well in showing trends chronologically and visualizes data changes at a glance. Line chart match best with periodical data but cannot be used to compare many categories in one line chart as it creates a mess.

- **Boxplot**

Box plots, also known as box-and-whisker plots, are a type of graph used to visualize the distribution of a dataset. In cricket, box plots can be used to visualize a variety of data, such as the runs scored by different batsmen, the number of wickets taken by different bowlers, or the results of different teams in a tournament.

A box plot consists of a box and whiskers, with the box representing the middle 50% of the data and the whiskers representing the rest of the data. The median (or middle) value of the data is shown by a line inside the box. Additionally, the minimum and maximum values of the data are shown by dots outside the whiskers.

Box plots are a useful tool for comparing the distribution of data between different groups.

For example, you could use a box plot to compare the runs scored by different batsmen in a team, or the wickets taken by different bowlers in a match. The box plot would show you the range, median, and other summary statistics of the data for each group, allowing you to quickly see how they compare.

3. Results

Visualizations and Findings

In this section we will explore the visualizations we have created as a part of this project. Each visualization answers a question that pertains to the game of cricket and provides us with actionable insights which could possibly be useful for cricket teams. One set of visualizations pertain to international cricket and the other set focuses on the IPL (Indian Premiere League) which is the most viewed cricket league in the world.

For each visualization type we describe the visualization and then explain the findings / insights we obtain from the visualizations following which we list out possible limitations of these visualizations.

3.1 The 'HOME' factor (International Cricket)

In this section we will visualize the effect of home/away games on a match (across various formats of the game for international teams) using a set of visualizations and try to visually answer the question "**Is home advantage a real thing? If so, how does it affect the cricket matches?**"

3.1.1 Home Factor for Winning in ODIs (One Day International games)

Fig 1. The home factor for winning in ODIs

⇒ **Description:**

The above visualization shows how the home factor influences the results of a match.
The Color gradient from red to green across shows the number of home matches played. Size of dots show the winning % of Home Matches. The marks labeled across each country depicts Home Matches, % of Home Matches Won and the country name in order.

⇒ **Findings:**

- The average scale of home matches hosted per country towards the east of Afghanistan are higher than the other side
- More than 60% of the matches hosted by South Africa, Pakistan, Australia are won in the home grounds, with Australia having a high advantage of home factor by winning 71% of matches hosted.

Fig 2. Win-Loss in Home/Away in ODI

⇒ **Description:**

The above chart shows the split of % of matches Lost/Tied/Won (indicated by color) for each country broken down by the location type where it's played w.r.t country into two sections: Home/Away.

The width of the bar chart shows the % of Total Matches across the result.

There is another indicator for the absolute number of total matches played by the respective country which is indicated by the mark on top of the respective country map icon to get a perspective of the scale of the matches played by any given country in Home/Away locations

⇒ **Findings:**

- Home advantage has played a crucial factor in gameplay for Australia, Bangladesh, New Zealand where the winning % has seen a major drop of greater than 20% of the games played respectively, from games played at home to the games played away from home

3.1.2 Home Factor for Winning in T20s (Twenty 20)

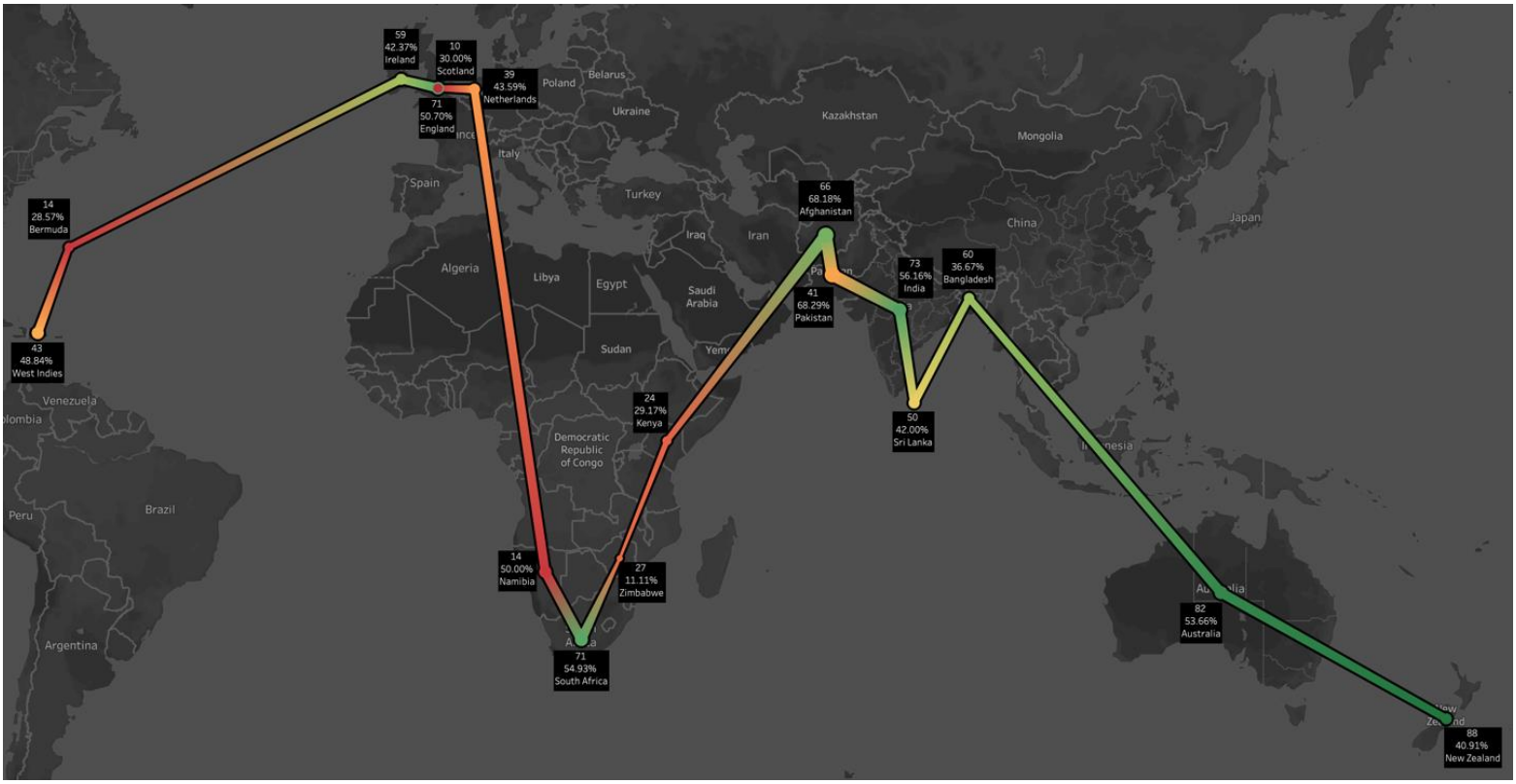


Fig 3. The home factor for winning in T20s

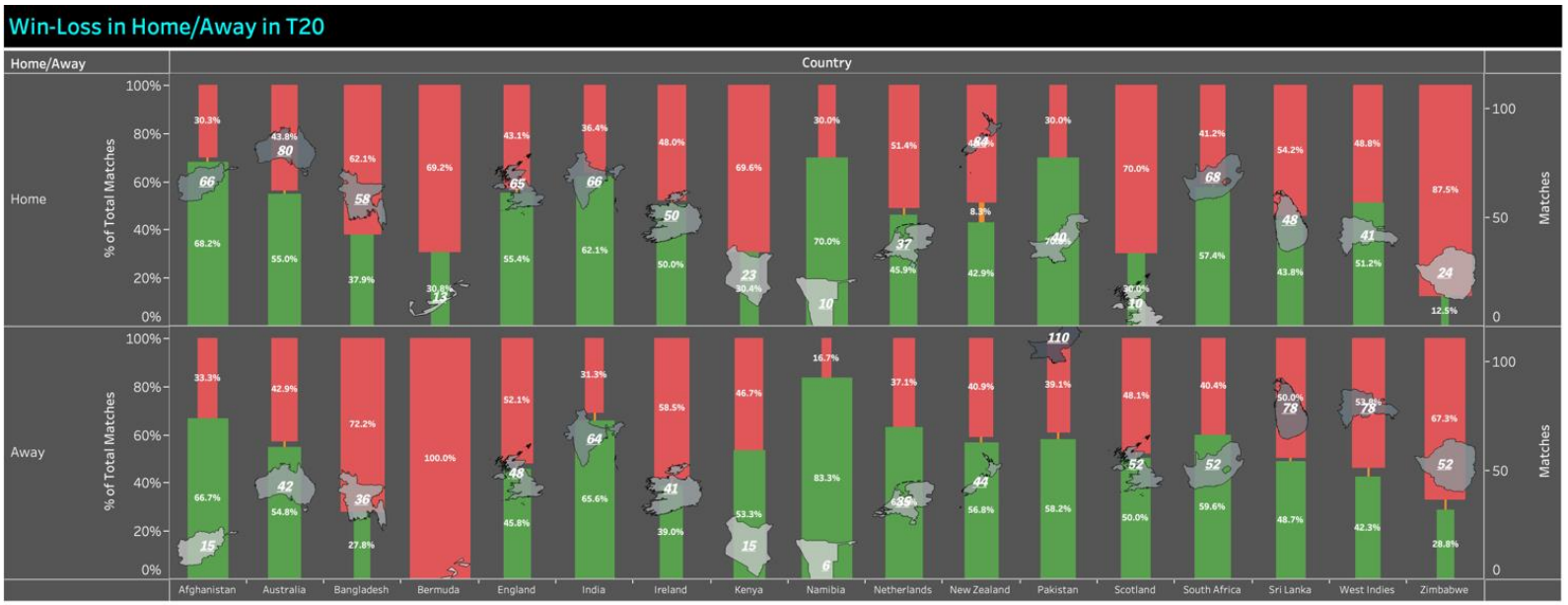
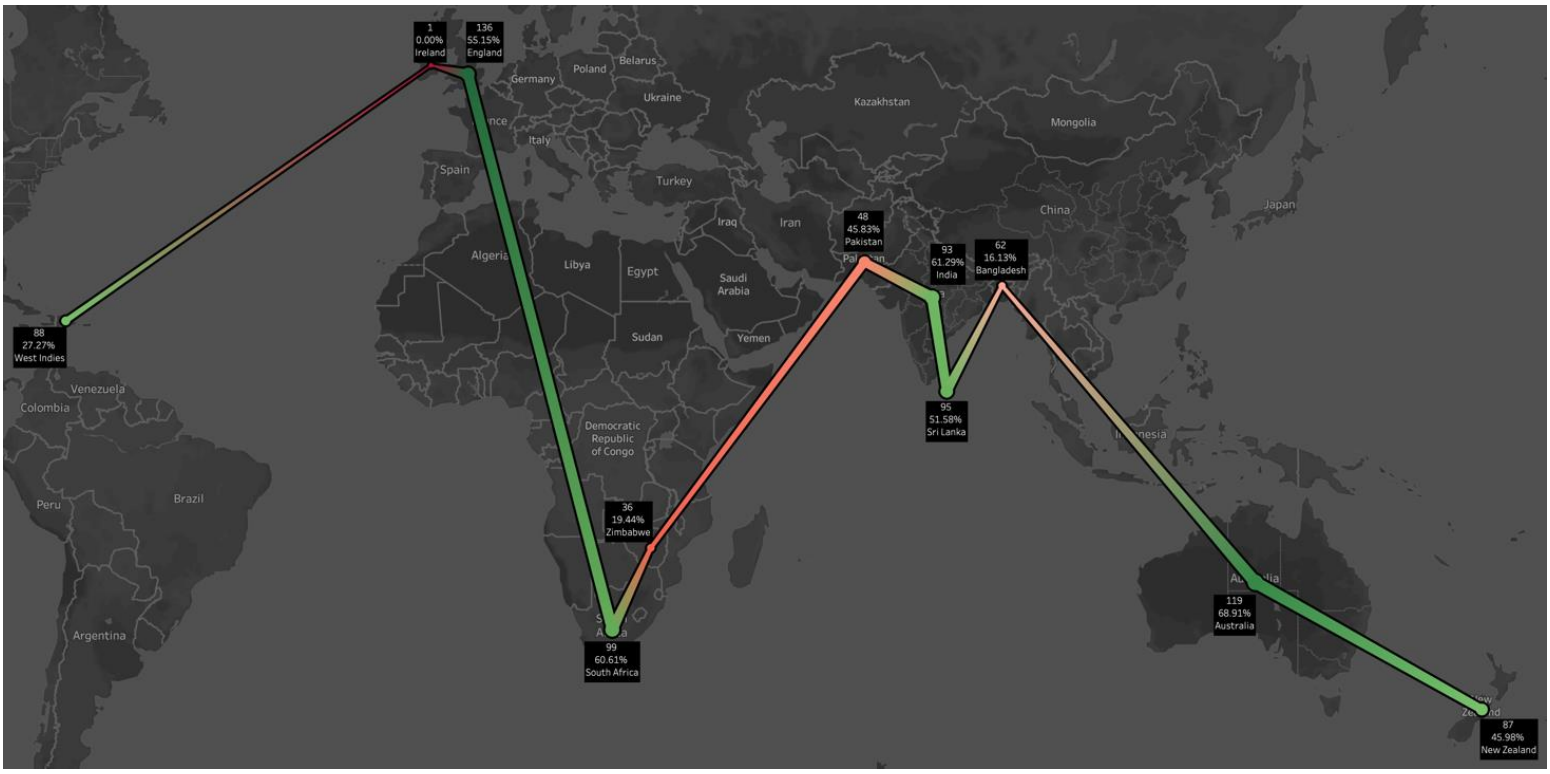


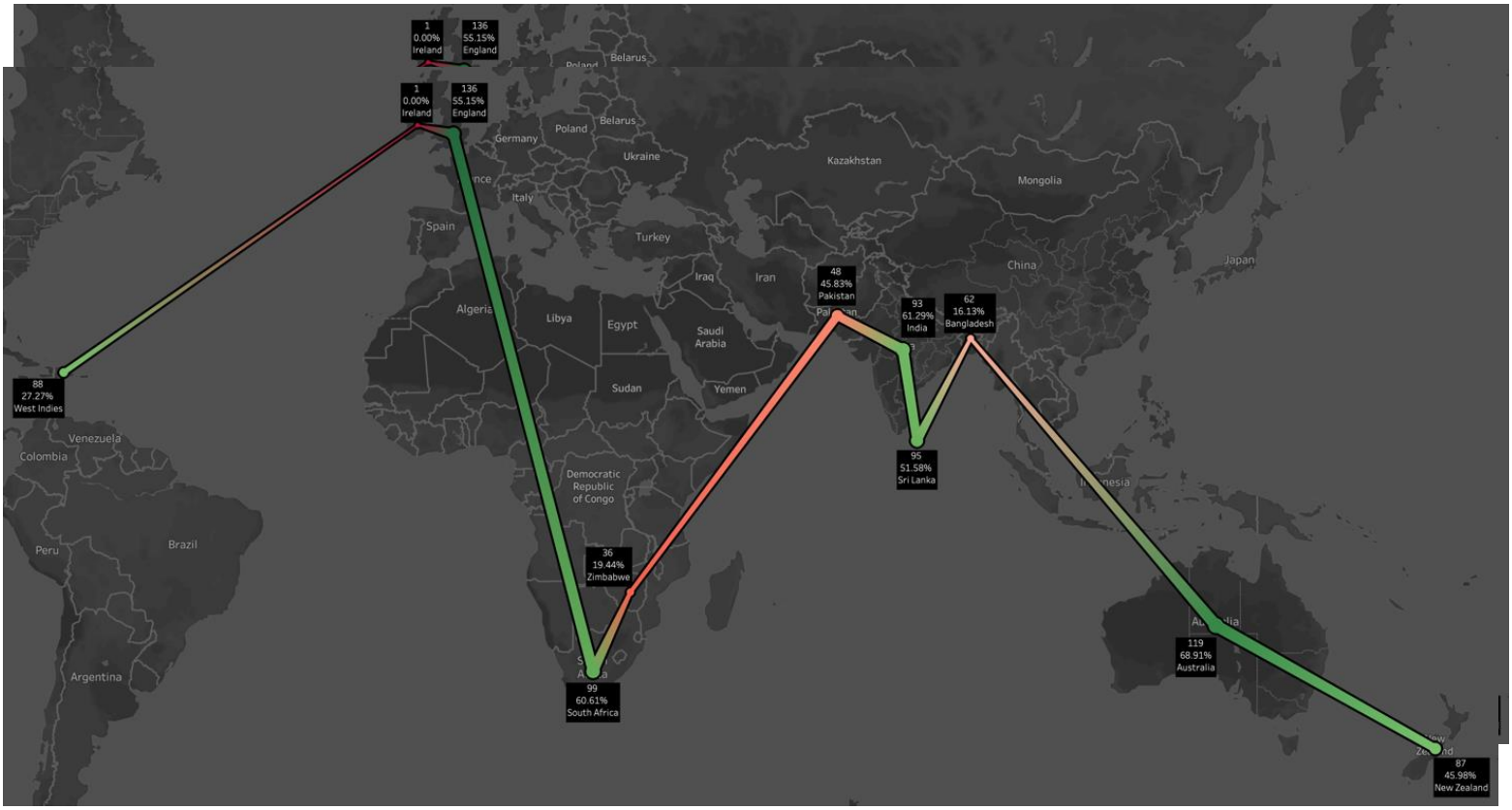
Fig 4. Win-Loss in Home/Away in ODI



⇒ **Findings:**

- The average scale of home matches hosted per country towards the east of Afghanistan are higher than the other side.
- Considering the diverse scale of matches played by the country in order to draw comparison, the ‘away’ factor worked in favor for India, Netherlands, New Zealand, South Africa, Sri Lanka and Zimbabwe rather than the ‘home’ factor. The above-mentioned teams have won higher proportion of matches while playing away from the home grounds rather than playing in home grounds.

3.1.3 Home Factor for Winning in Test



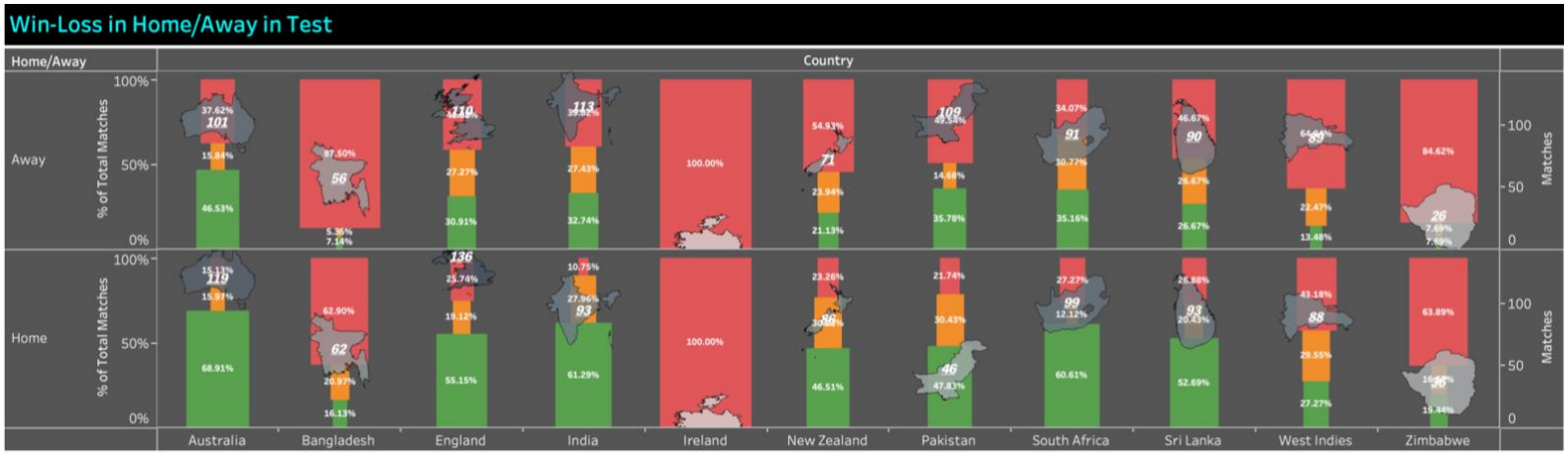


Fig 6. Win-Loss in Home/Away in Test matches

⇒ **Findings:**

- Given the scale of test matches hosted, England, Australia and New Zealand have hosted the highest number of matches in the 21st century until now
- While all the teams mentioned in the visualization have an advantage of the home factor, teams like India, New Zealand, South Africa and Sri Lanka had distinctively evident home advantage, where the winning % has seen a major drop of greater than 25% of the games played respectively, from games played at home to the games played away from home

3.1.4 Home Factor in Win margin

Here we look at a slightly different visualization that shows us win margins of teams versus other teams when they play home or away, this helps answer the question of **how being home or away affects a team's performance and the margin of wins** they have when they are home or away.

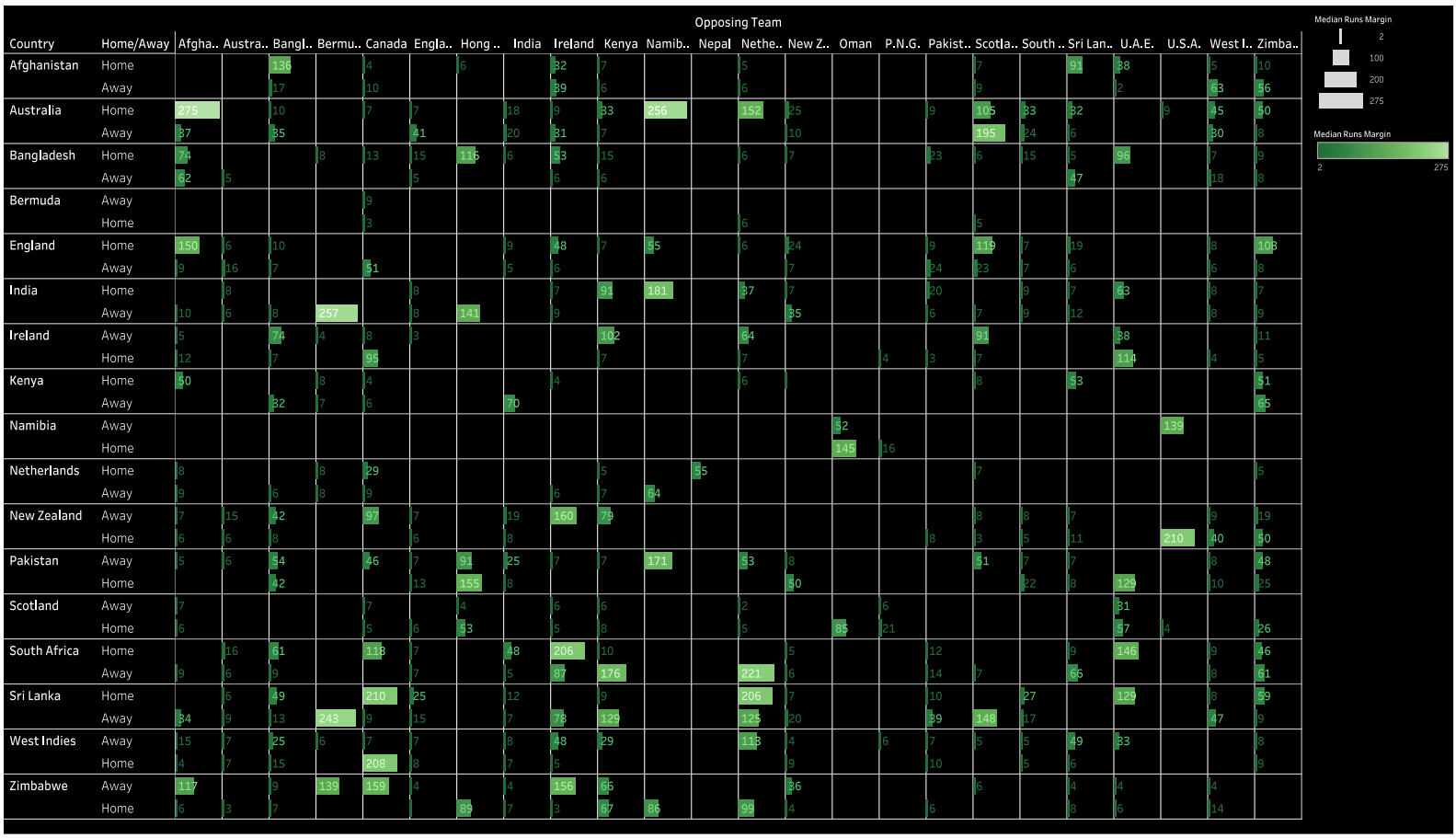


Fig 7. Median Runs Margin for teams in ODIs

⇒ Description:

This visualization above shows the runs margin of each team's (mentioned in the left side band) median runs margin while winning against the opposing team split by Home/Away location

⇒ Findings:

- India (vs Bermuda), New Zealand(vs USA), Sri Lanka(vs Bermuda), West Indies(vs Canada) teams have their highest median runs margin while winning in matches away from home rather than in home matches
- Teams have scored higher margins (>100 runs) with higher frequency (>3) against teams like Canada, Netherlands, Scotland, UAE

3.2 The Rise of cricket (Across all forms of cricket)

We look at the number of matches played on an annual basis across all the leagues in cricket to understand the rise in popularity of cricket by creating a simple bar plot for the same as seen below.

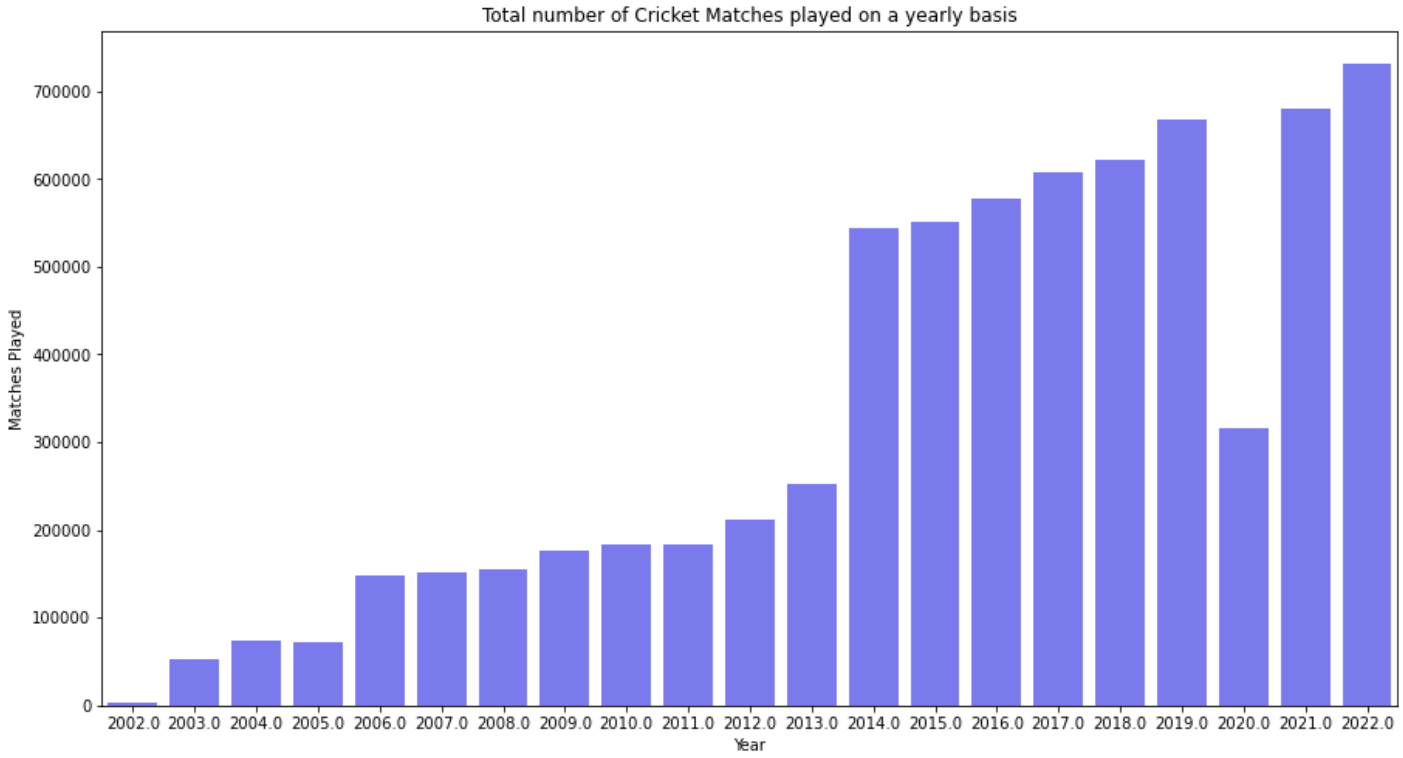


Fig 8. Total number of Cricket matches played annually

⇒ **Findings:**

- We see a steady rise in the number of matches thus implying a rise in the popularity of cricket as a sport across the world in the past 2 decades.
- There is an exception in the year 2020 however, this is due to the COVID-19 pandemic which drastically reduced the number of games being played across the world.

3.3 Overall scoring trends (IPL)

In this section we observe some general trends and try to visualize some overall trends in the IPL.

⇒ **Description:**

We now explore the IPL data which is a subset of the entire cricsheet dataset.

The first visualization (Fig 9) is a bar chart that shows us the number of matches being played.

The second visualization (Fig 10) is a line chart that shows run totals on a yearly basis.

The third visualization (Fig 11) is a bar chart that portrays boundaries on a yearly basis.

The fourth and final visualization (Fig 12) shows the number of teams participating in the IPL league each year.

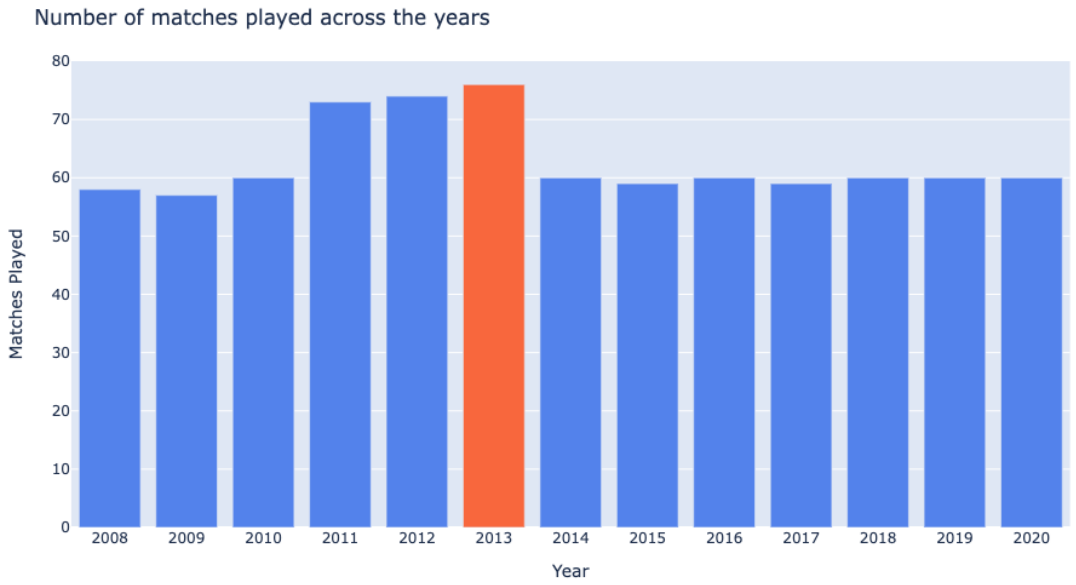


Fig 9. Number of matches played across years

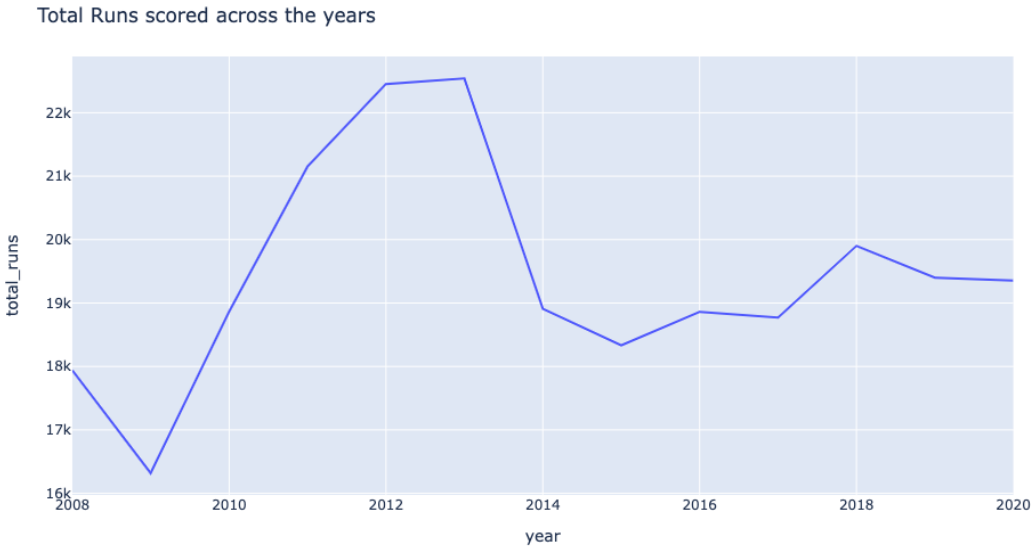


Fig 10. Total runs scored across years

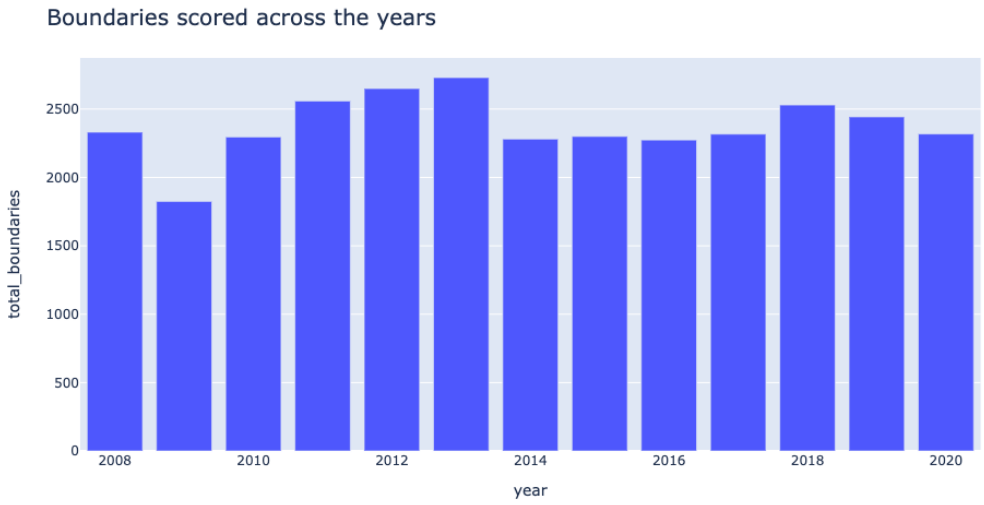


Fig 11. Number of Boundaries scored across years

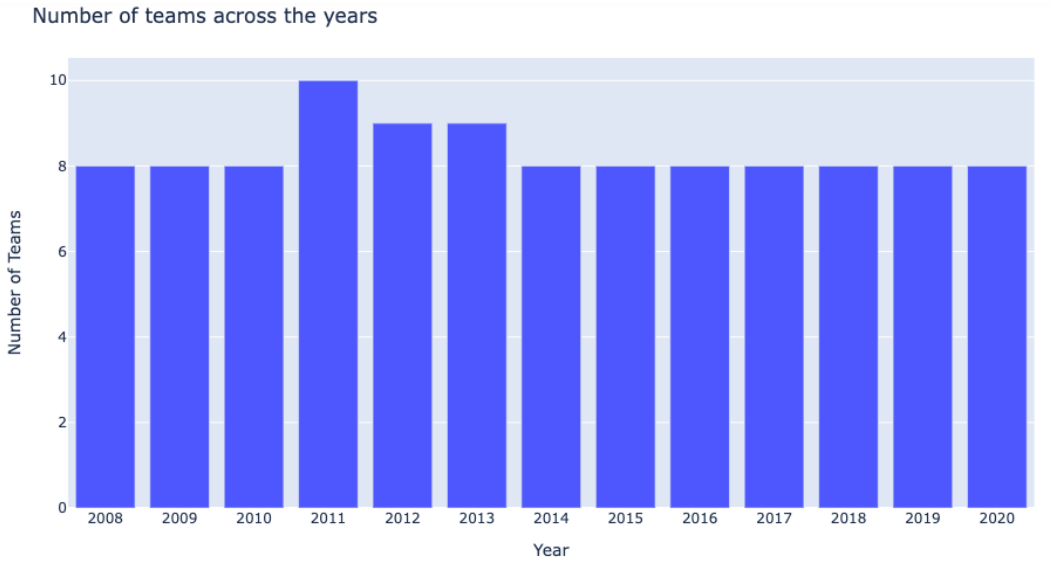


Fig 12. Number of teams participating across years

⇒ **Findings:**

- On observing the visualizations Fig 9 and Fig 10, we see that the number of matches, runs scored increase from 2011 – 2013, looking at Fig 12. we can see that the only reason for this is that there is an increase in the number of teams in those years.
- Fig 10 and Fig 11 seem to mirror each other in terms of trends thus implying that boundaries play a fundamental role in scoring runs. This makes sense as boundaries are worth 4 or 6 runs as opposed to the usual 1 or 2 runs in a game.
- Apart from 2011 - 2013, there is no other variation in the number of games played as IPL is a 2-month long league. It is however interesting to note that there is increase in the number of runs scored in 2014 -2020 despite having the same number of teams.

3.4 Team level scoring trends (IPL)

⇒ **Description:**

In this section we look at 2 visualizations that show runs scored and conceded respectively at a team level.

The first visualization (Fig 13) is a line chart that shows the average runs scored by a batting team in a given year, the x axis represents the year, and the y axis represents the average runs scored by the batting team.

The second visualization (Fig 14) is similar to the first one except over here the line chart shows runs conceded by a bowling team in a given year. The x axis represents the year, and the y axis represents the average runs conceded by the bowling team.

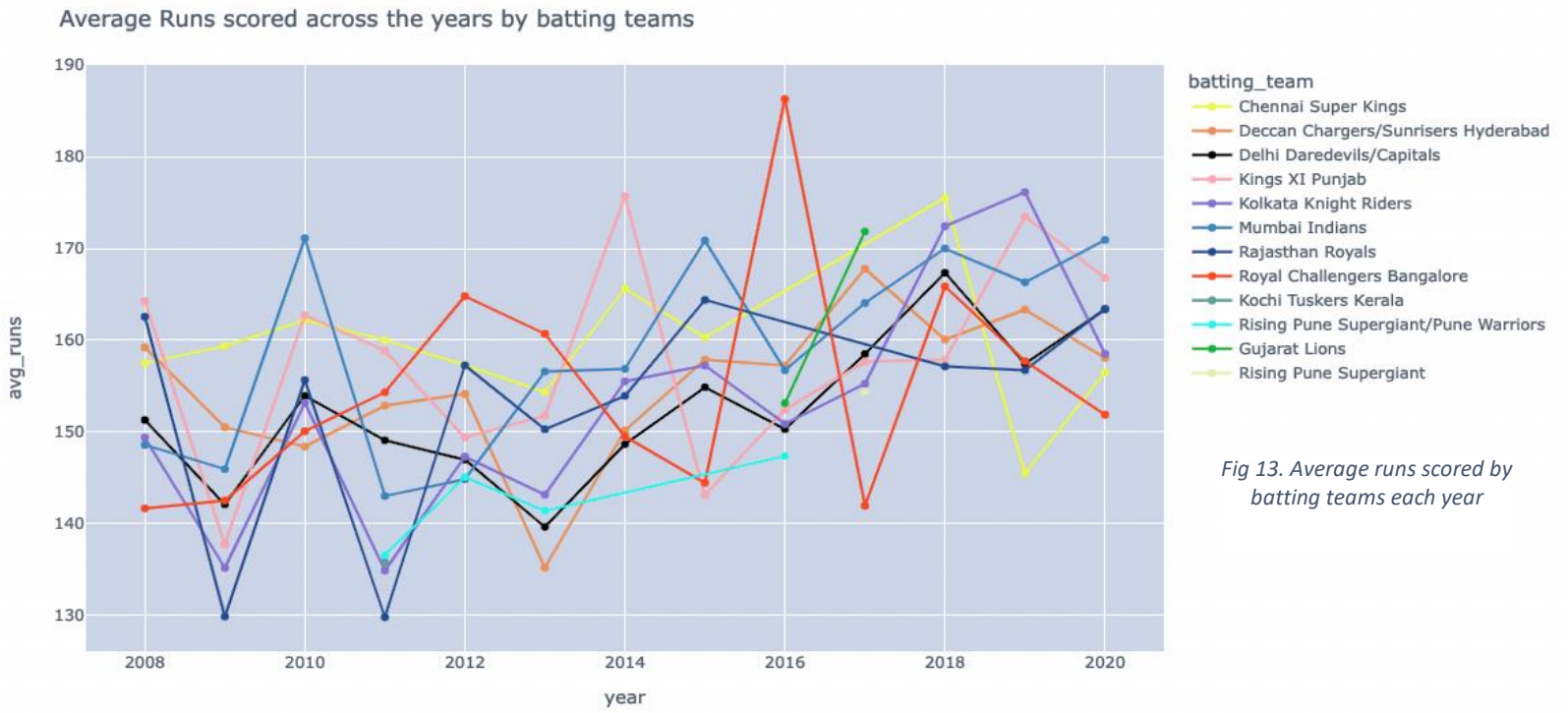


Fig 13. Average runs scored by batting teams each year

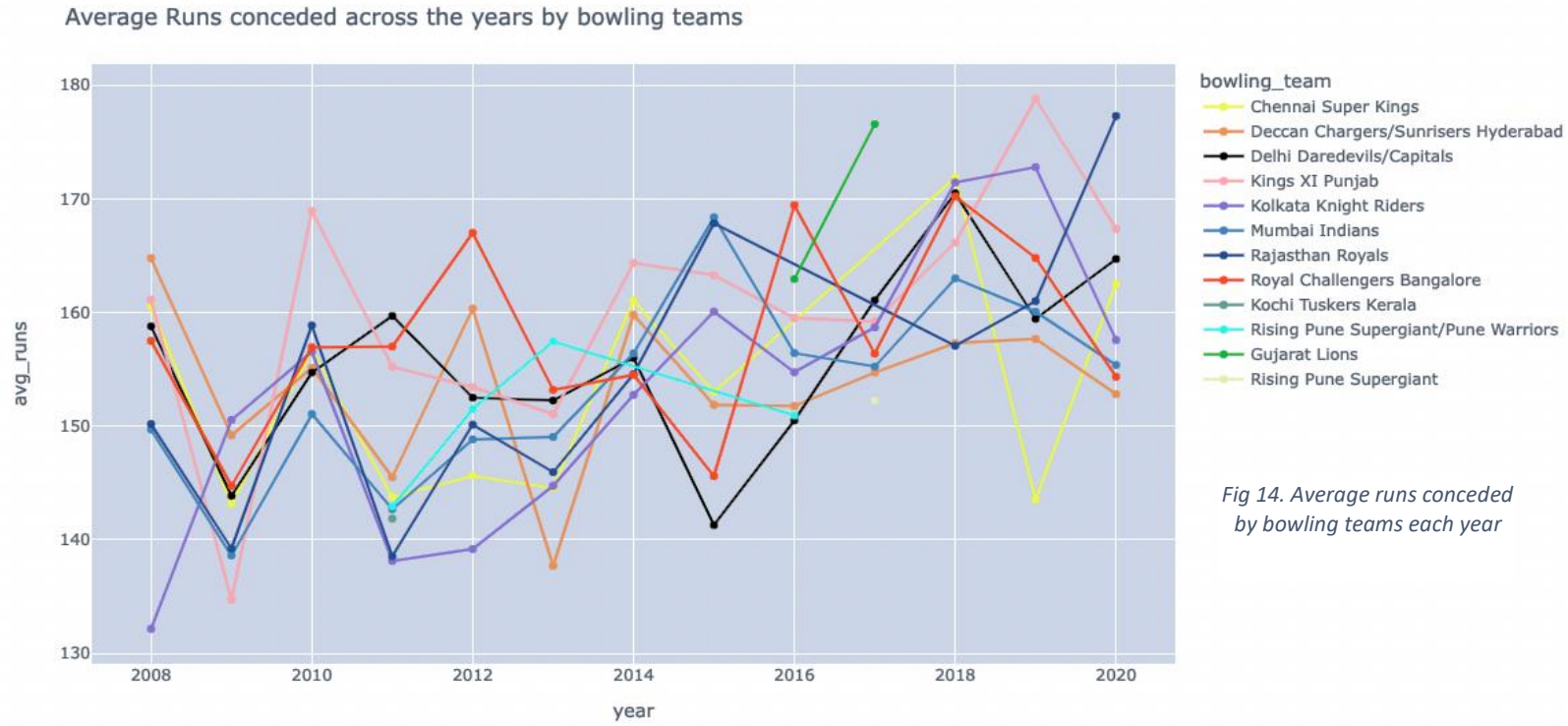


Fig 14. Average runs conceded by bowling teams each year

⇒ **Findings:**

- On observing both the figures above it is evident that teams that score more runs, concede less runs on an average are usually the more successful ones.
- For instance, let us look at the year 2020, the team Mumbai Indians (blue) scored the highest runs and conceded the least runs and as expected were the winners of the IPL that year.
- Another instance is in the year 2010, the team Mumbai Indians (blue) scored the highest runs and conceded the least runs and as expected did very well and made it to the finals. They were runner ups behind the team Chennai Super Kings (bright yellow) who were closely tailing Mumbai in both runs scored & conceded.
- Scoring higher runs and conceding less runs is not the only way to win eg. in the year 2018 the team Chennai Super Kings won but had the highest runs conceded, to combat this they also scored highest runs that year.

3.5 Is the toss a deciding factor in the team's performance? (IPL)

In this section we attempt to understand the effect of a toss on a team's performance. Like most other games, teams decide to choose how they start (batting or fielding/bowling) via a coin toss. We examine these factors and try to understand what the best decision is to make at the time of a toss.

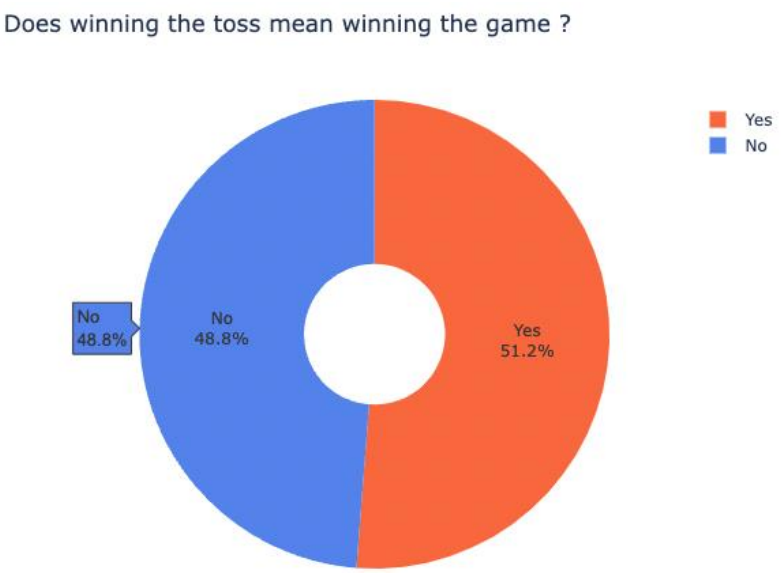


Fig 15a. Does a toss win guarantee a game win?

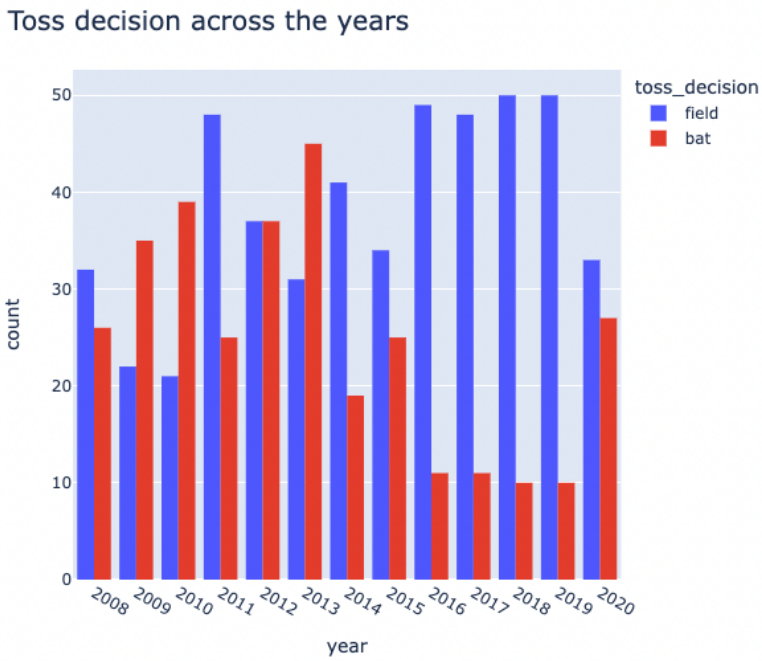


Fig 16. Toss decision counts each year

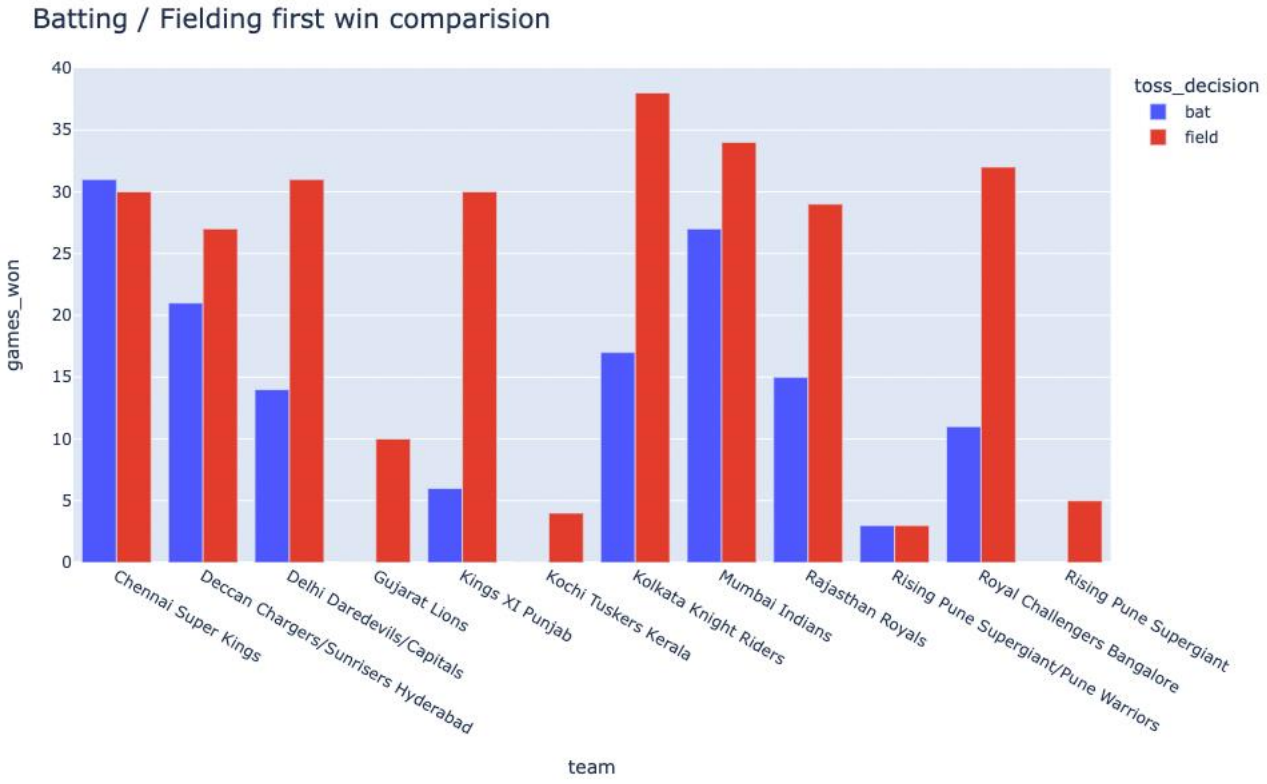


Fig 17. Batting or fielding win breakdown for each team

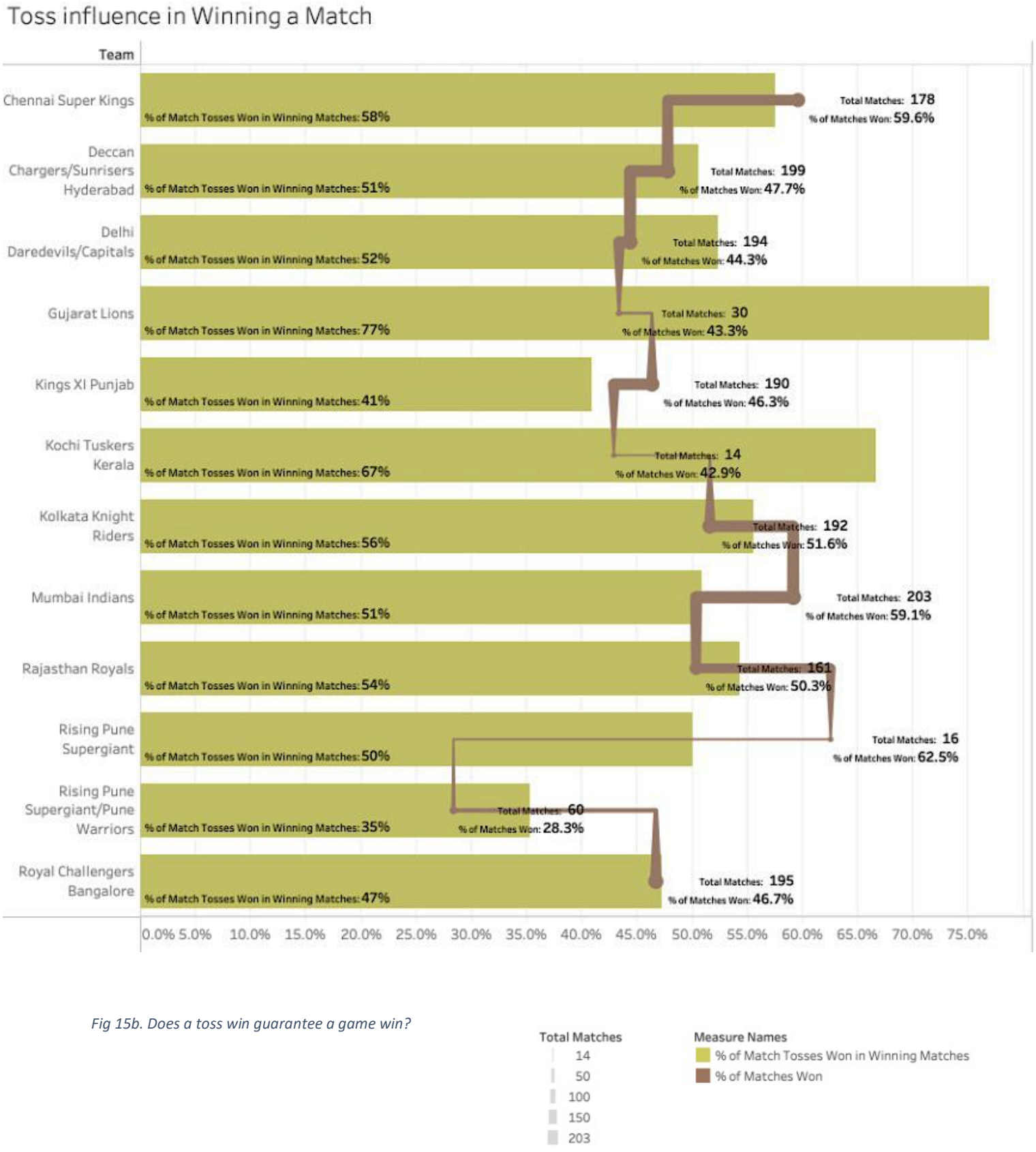


Fig 15b. Does a toss win guarantee a game win?

⇒ **Description:**

- In Fig 15a we check the win status of teams that have won a toss and create a pie/doughnut chart to see the breakdown of wins / losses for teams that have won the toss.
- In Fig 15b we create a bar chart that shows the influence of the winning toss on winning matches in IPL. Here % of match tosses won in winning matches is represented by bar graph and the % of Matches won is represented by the point comparison graph and the scale of total matches played is represented by the thickness of the point
- In Fig 16 we use a stacked bar chart to count the frequency of decisions (batting or fielding/bowling) that teams choose at the toss. The x axis is the year, and the y axis is the count of tosses.
- In Fig 17 we use a stacked bar chart to visualize the toss decision breakdown amongst games that have been won by specific teams. The x axis represents the team, and the y axis represents the number of wins

⇒ **Findings:**

- Through Fig 15a we see that there is an almost 50-50 split between winning or losing a match even after winning the toss, almost rendering winning the toss useless.
- In Fig 15b we can see that Gujarat Lions is one of the luckiest team that won nearly 77% of matches in which it had won the tosses, although this could be because they have played only 30 matches which is comparatively less.
- In Fig 16 we see an interesting trend from 2009 – 2013 where teams opted to bat in favor of fielding, however from 2014 onward this trend is flipped, and teams are more in favor of fielding first.
- Fig 17 may throw some light on the shift in trend from 2014 onward, it seems that teams realized that fielding first is favorable in order to win the game as most teams have a better win percentage by fielding first.

3.6 The effect of Boundaries on a team's performance? (IPL)

In this section we look at the effect of boundaries on a team's performance and visualize overall trends with boundaries.

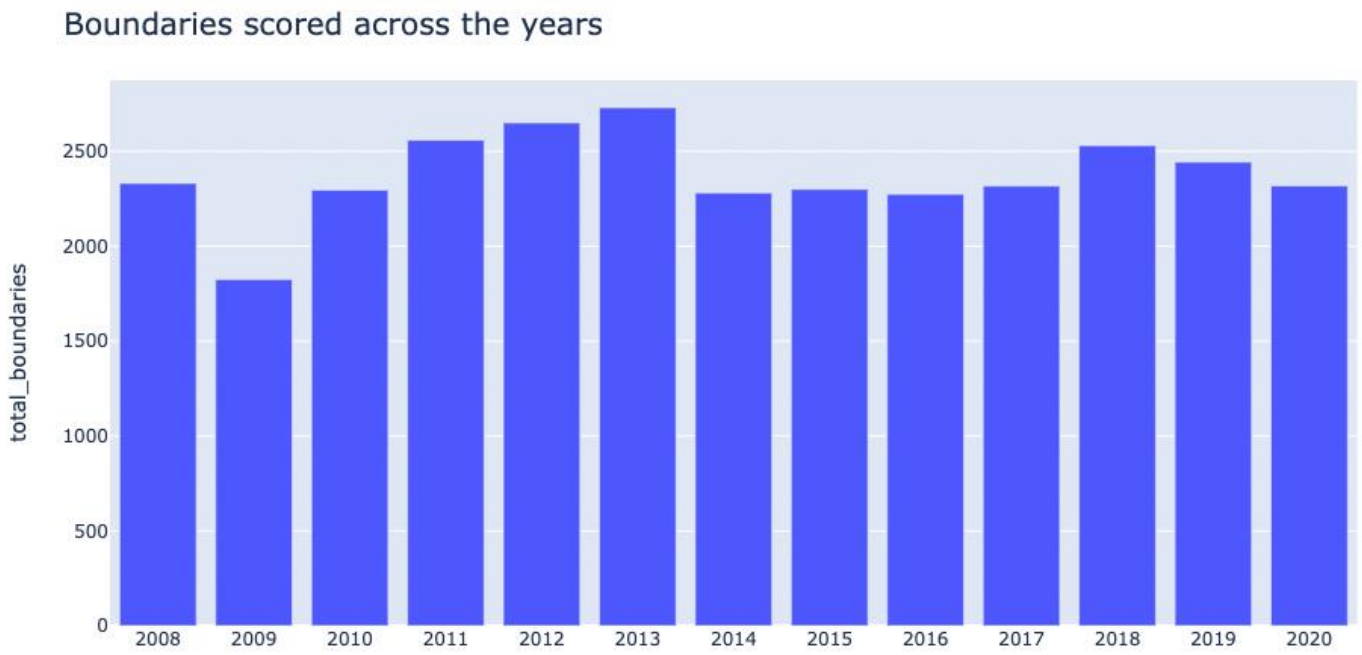


Fig 18. Boundaries scored each year in the IPL

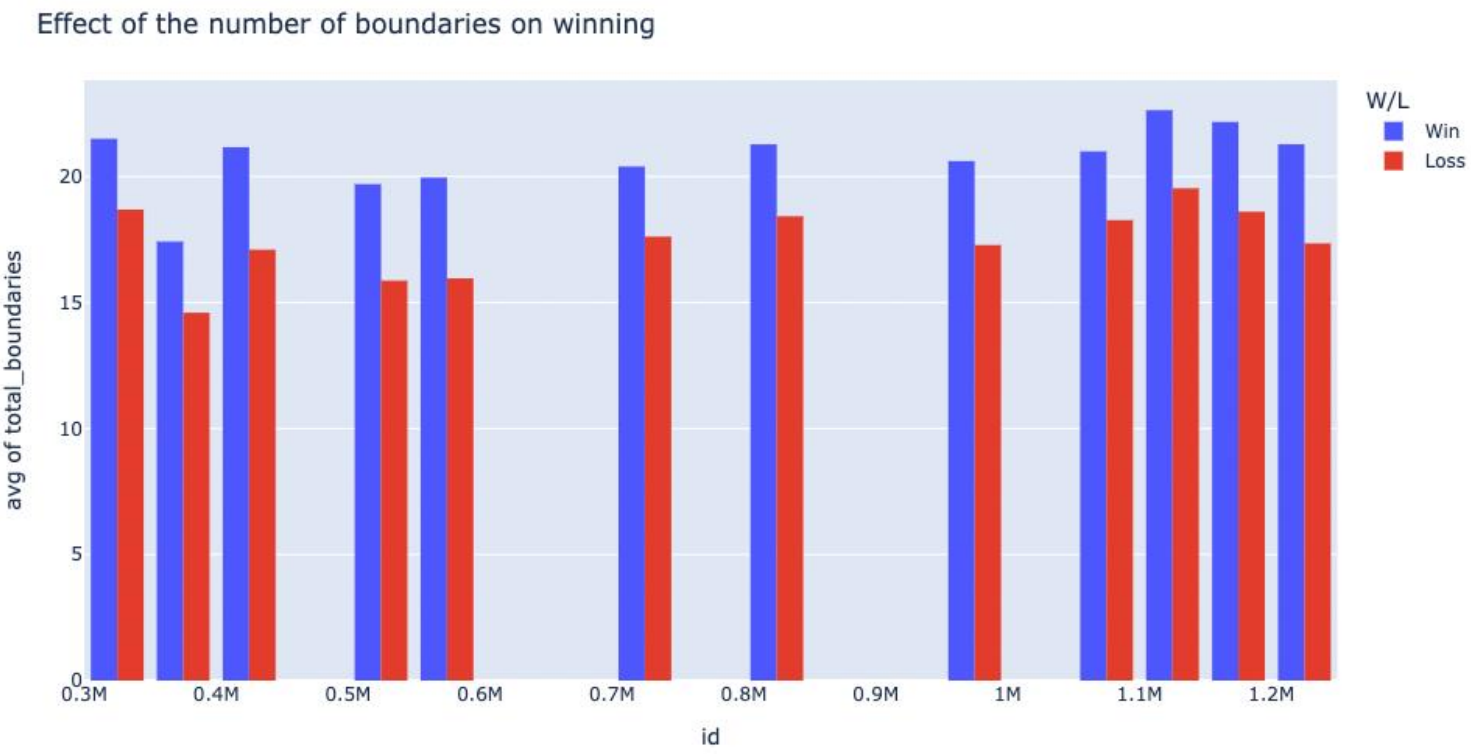


Fig 19. Average number of boundaries scored by teams (binned by ID) and win/loss breakdown

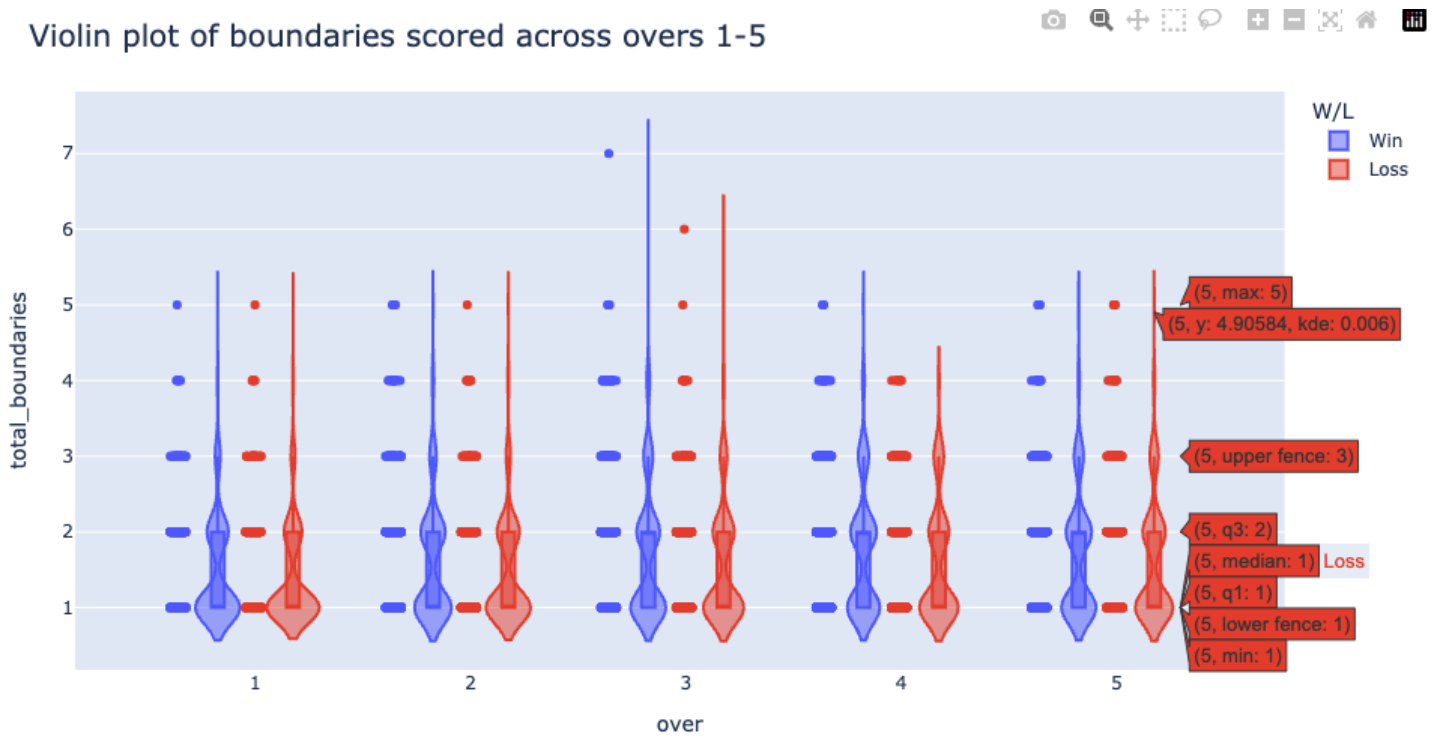


Fig 20. Distribution of boundaries scored across overs 1-5

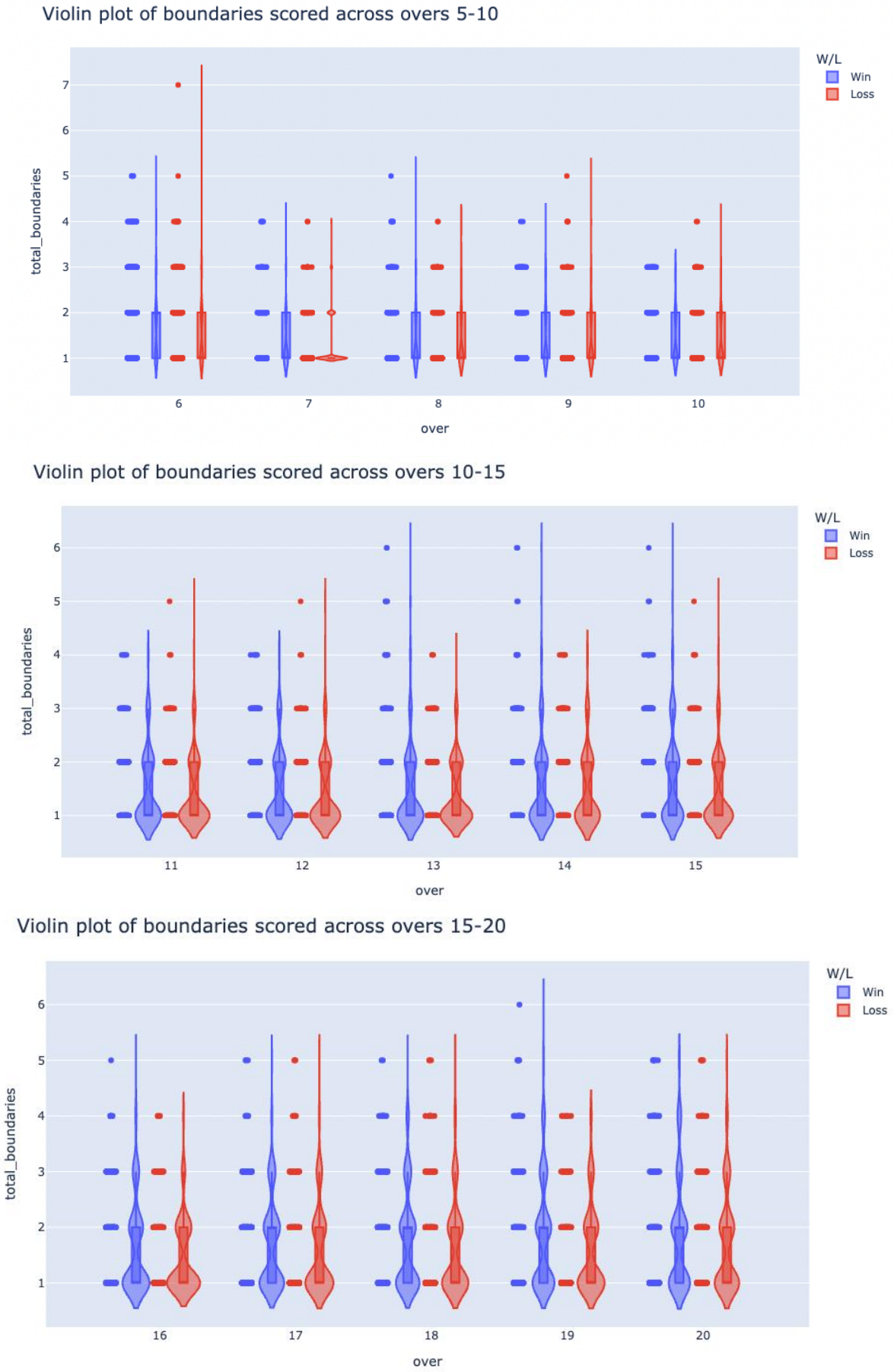


Fig 21. Distribution of boundaries scored across overs 5 – 20

⇒ **Description:**

- In Fig 19 we use a bar chart to summarize the number of boundaries each year of the IPL.
- In Fig 20 and 21 we use a violin plot to visualize the distribution of boundaries across each over of the game in winning vs losing teams.

⇒ **Findings:**

- In Fig 19 we see that the number of boundaries is almost consistent (with respect to the number of games each year) across the years with about 10% increase in the latter years indicating that players have become better run scorers.
- Through Fig 20 and 21 we see that between winning and losing teams the number of boundaries are mostly similar with the winning teams usually having a few outliers in terms of the number of boundaries.
- Another interesting to note is that from overs 6 – 10 there are a lot lesser number of boundaries overall.

3.7 The effect of Scoring rate on game outcome (IPL)

In this section we look at the scoring rate in the game to get an idea of team scoring patterns and also contrast winning versus losing teams in their scoring trends across overs.

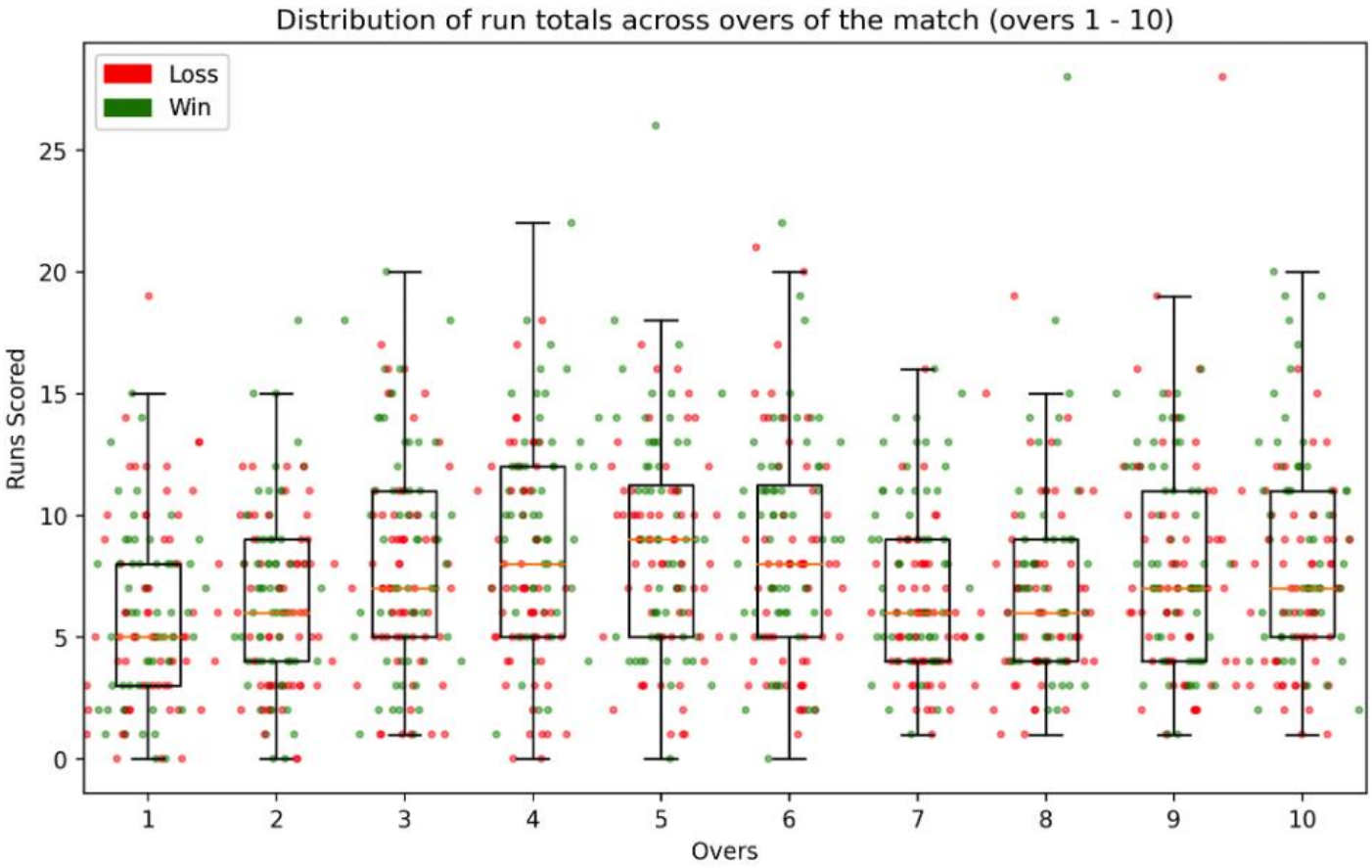


Fig 22. Distribution of runs scored across overs 1-10 broken down by win or loss.

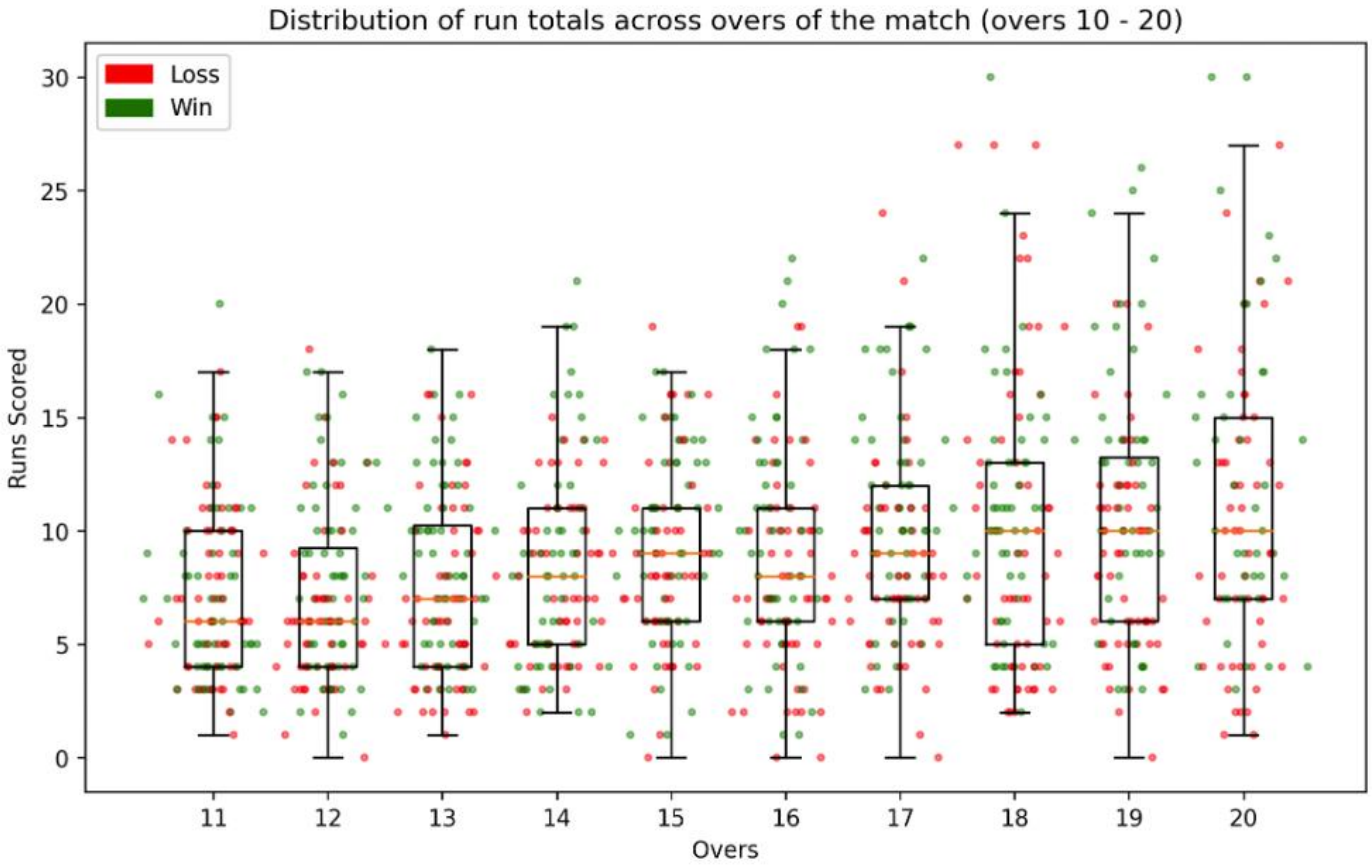


Fig 23. Distribution of runs scored across overs 10-20 broken down by win or loss.

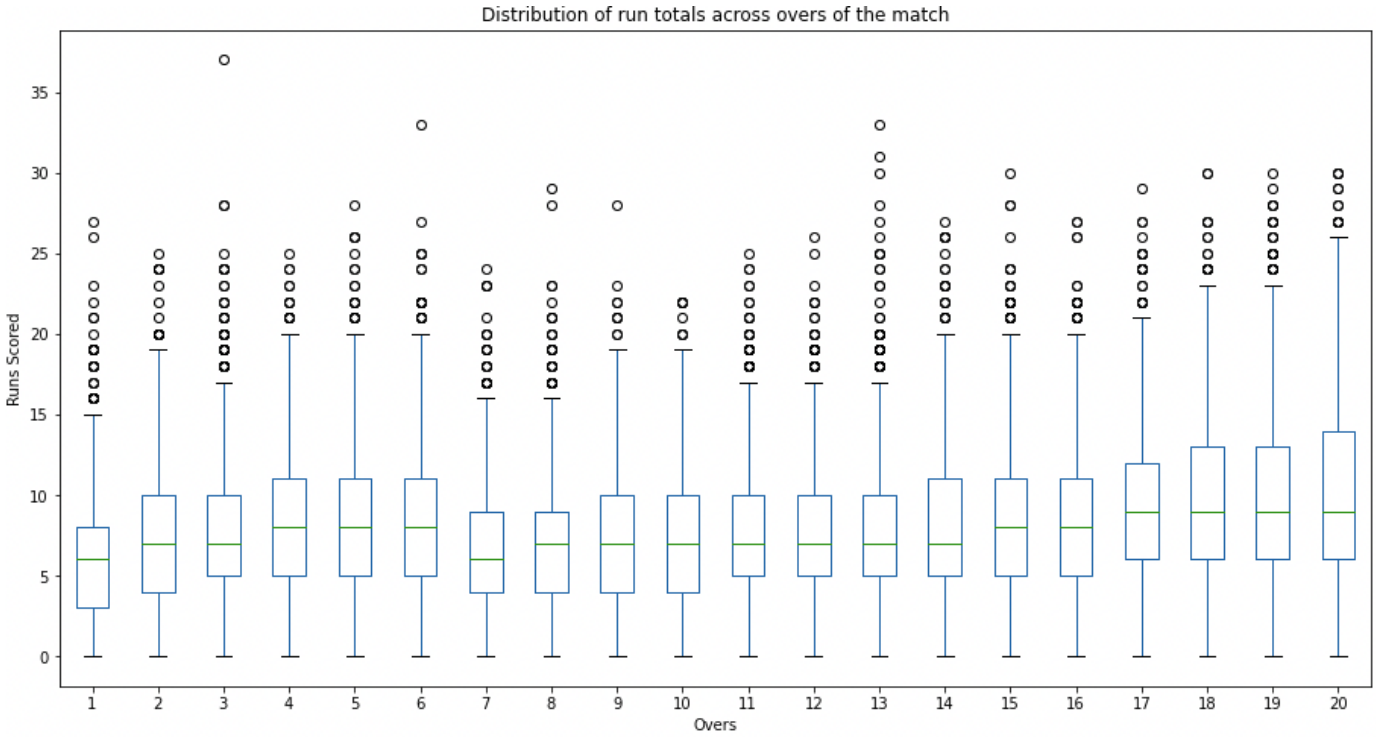


Fig 24. Distribution of runs scored across overs.

⇒ **Description:**

- In Fig 22 and 23 we use a boxplot with a scatter overlay to show the summary stats and distribution of datapoints that signify the runs scored in different overs.
- In Fig 24 we plot a regular boxplot to summarize the distribution of runs scored across the overs.

⇒ **Findings:**

- In Fig 22 we see that in general winning teams score runs that are after the 3rd Quartile and losing teams seem to be more concentrated around the median and quartiles below that. We also see a spike in runs scored in the 3rd and 4th overs.
- In Fig 23 we notice a similar trends to Fig 22 with a spike in the number of runs scored in overs 18 – 20.
- Fig 24 also shows the same trends as figures 22 and 23 in a more succinct manner.

3.8 The effect of venue on scoring (IPL)

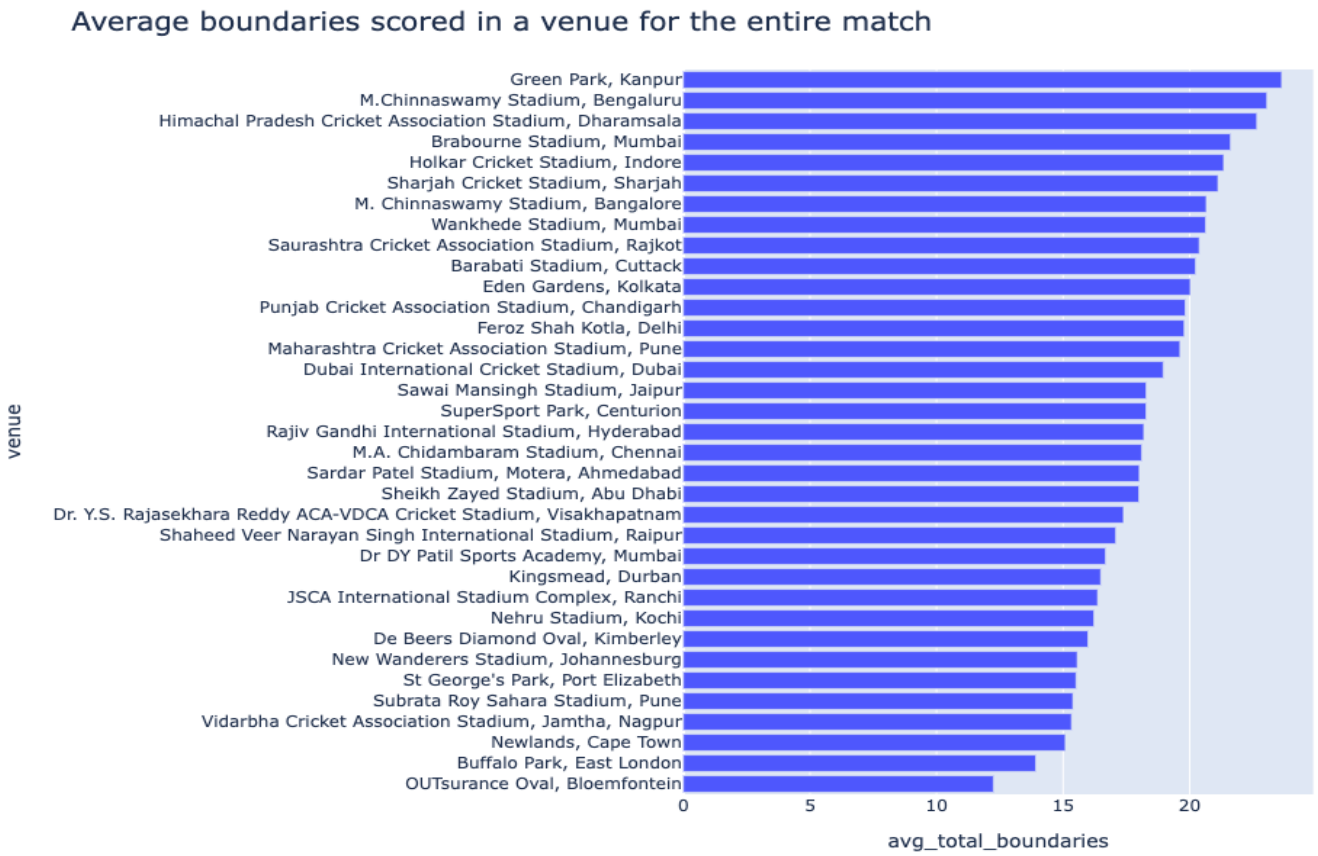


Fig 25. Average number of boundaries scored locationwise

⇒ **Description:**

- Fig 25 shows us the average number of boundaries scored at a given location.

⇒ **Findings:**

- Kanpur, Bangalore and Dharmasala have high scoring stadiums. This could possible explain why the team Royal Challengers Bangalore had high average conceded runs and high average runs scored (Fig 13 & 14) as the High scoring M. Chinnaswamy Stadium is their home stadium.

3.9 Player Comparision

'King' Kohli and *'Master Blaster'* Sachin Tendulkar have been regarded as the greatest batsmen till date to wear the Indian team jersey.

Despite Sachin retiring from cricket, he is still compared to many of the talents currently playing for the Indian Blues. Moreover, Sachin is mainly compared to Virat Kohli as he is most likely to break the Master Blaster's records. In this analysis, we shall dive into the statistics of Virat Kohli vs Sachin Tendulkar in International cricket formats.

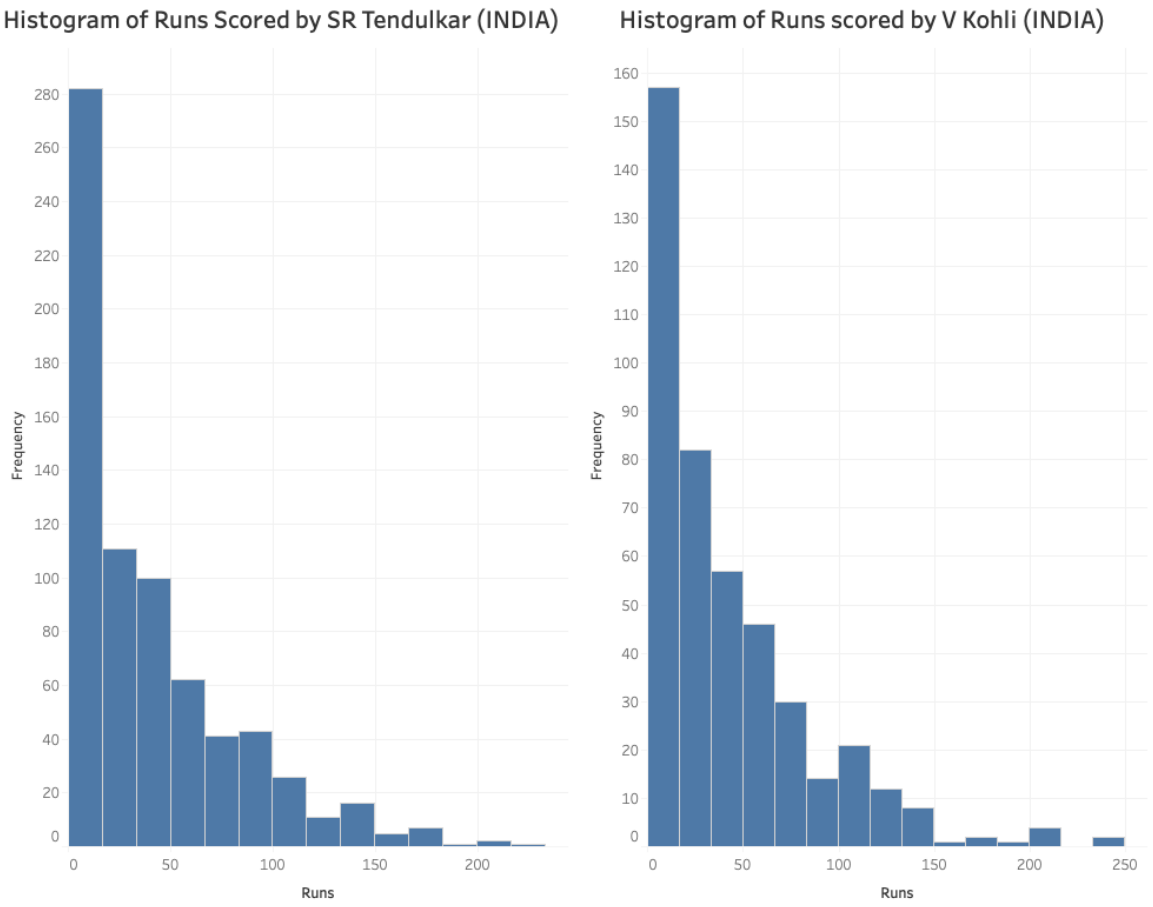


Fig 26. Histogram of runs scored by both players in individual games

⇒ **Description :**

The above histograms tell us about the range over which Virat Kohli and Sachin Tendulkar have scored runs as well its frequency.

⇒ **Findings**

- Both Virat and Sachin have mostly scored runs in the range of 0–50.
- The percentage of centuries scored by Sachin is far more than that of Virat.
- The number of games played by Sachin is far higher than that of Virat indicating that Sachin's career is much longer than Virat's career.

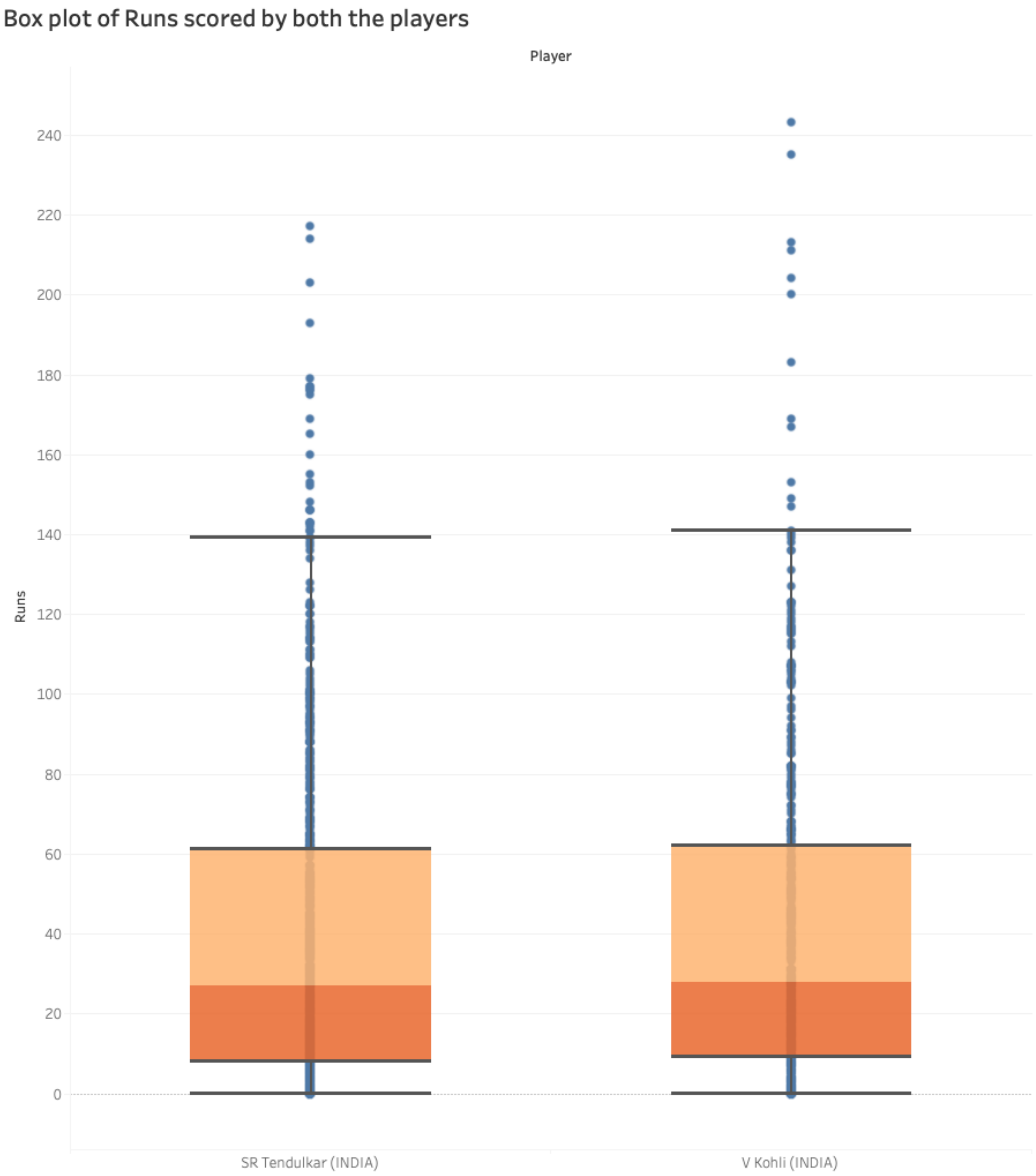


Fig 27. Boxplot of runs scored by both players in individual games

⇒ Description

The above boxplots not only tell us about the range of the runs scored but also emphasize on the median of the runs scored by both.

⇒ Findings

- The median score for Virat is 28.
- The median score for Sachin is 27.

Also, the extreme lines in the figures known as whiskers indicate the least and the highest runs scored by both. The least runs scored by both is zero and the maximum is 243 for Kohli and 217 for Sachin.

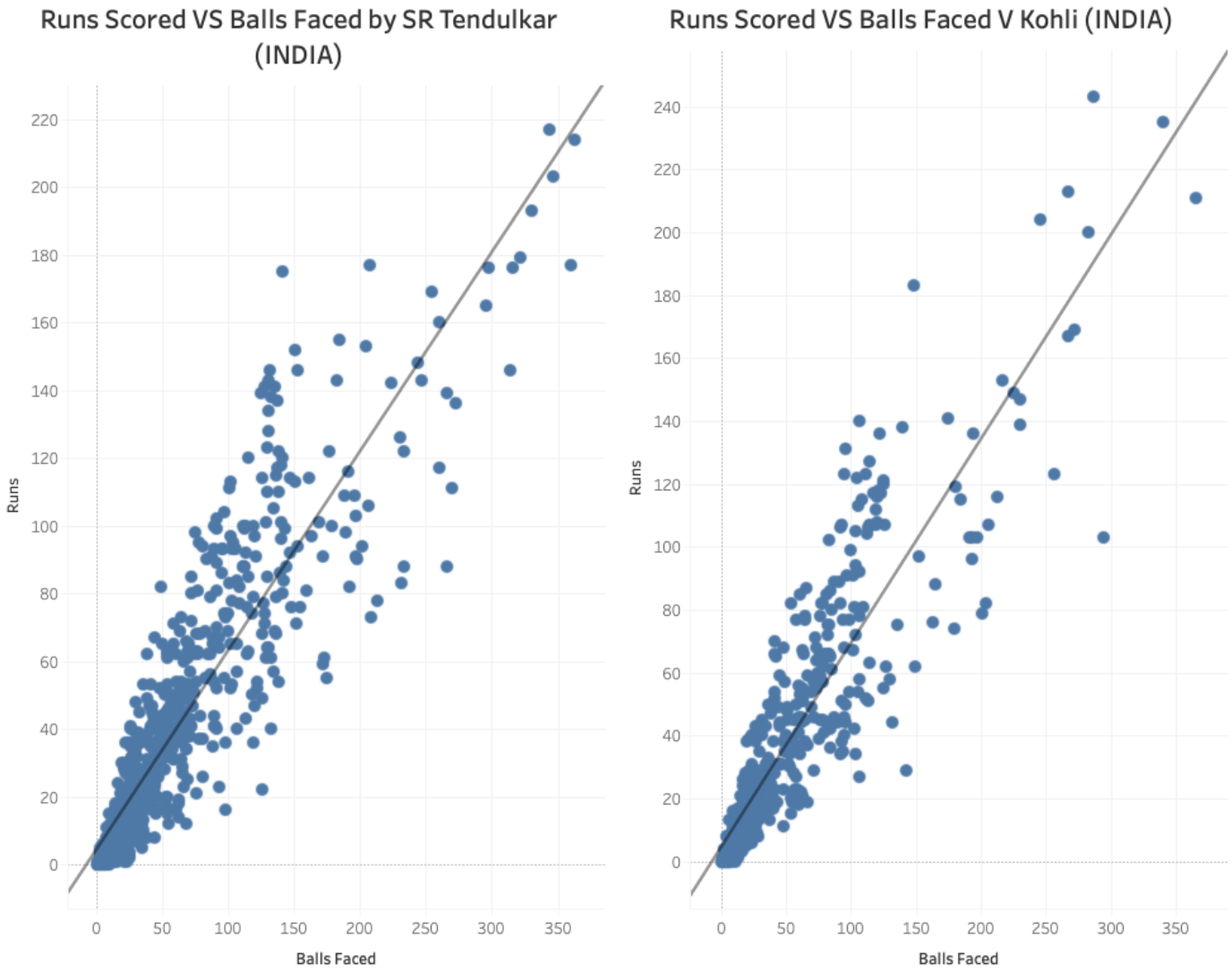


Fig 28. Runs scored vs balls faced by both players

⇒ **Description**

These visualizations show relation between balls faced and number of runs scored. Both the scatter plots have been fitted with a linear regression line.

⇒ **Findings**

- Virat's plot has a higher slope than that of scahin's, which means that for every extra ball faced by virat we can expect the runs to increase on average by a much higher amount than Sachin.

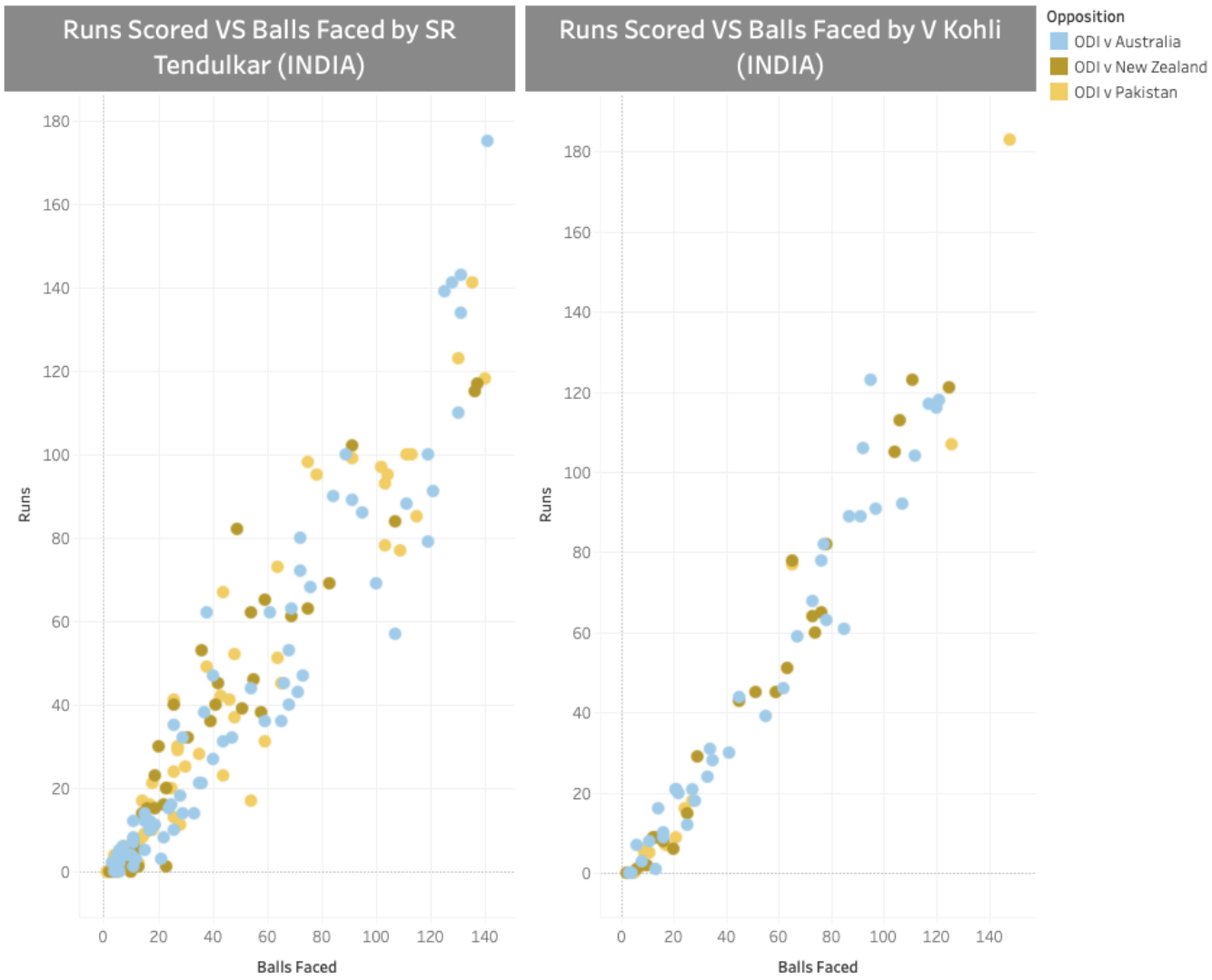


Fig 29. Runs scored vs balls faced by both players against specific teams

⇒ **Description**

These Visualizations show relationship between balls faced and number of runs scored when the opponents are Australia, New Zealand, and Pakistan in an ODI.

⇒ **Findings**

- Overall, the runs scored by Tendulkar is higher even when number of balls faced is small as compared to Kohli where there is nearly a linear relationship between balls faced and runs scored.

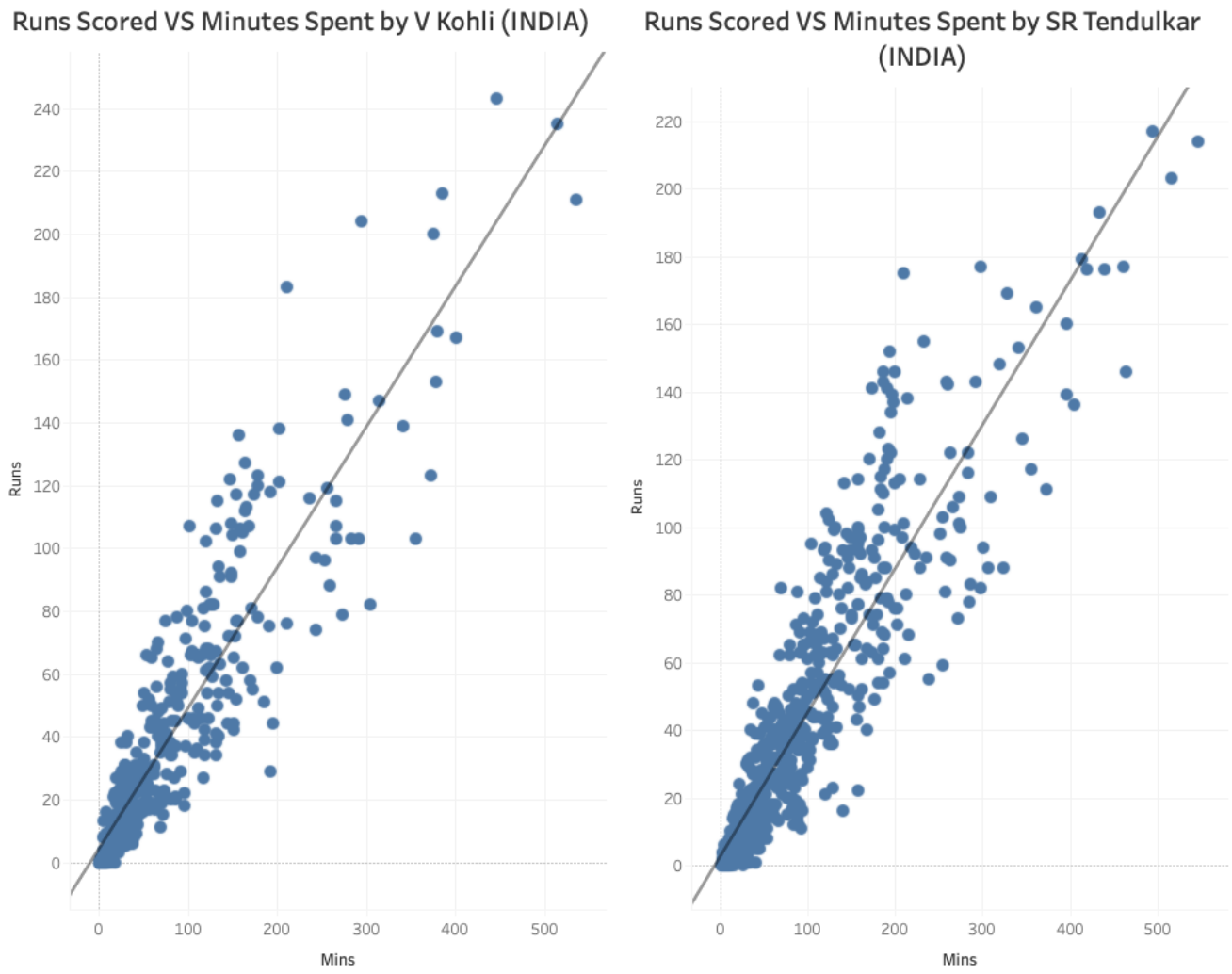


Fig 30. Runs scored vs minutes played by both players

⇒ Description

These visualizations show relationship between minutes on field and runs scored.

⇒ Findings

- We can see that Virat Kohli makes more runs in lesser time compared to Sachin Tendulkar.
- The gap is initially less where in at 100 mins, we can see that Sachin is just a few runs less compared to Virat. However, this gap becomes more as minutes on field increases, wherein Virat is able to hit more runs compared to Sachin as minutes on field increases.

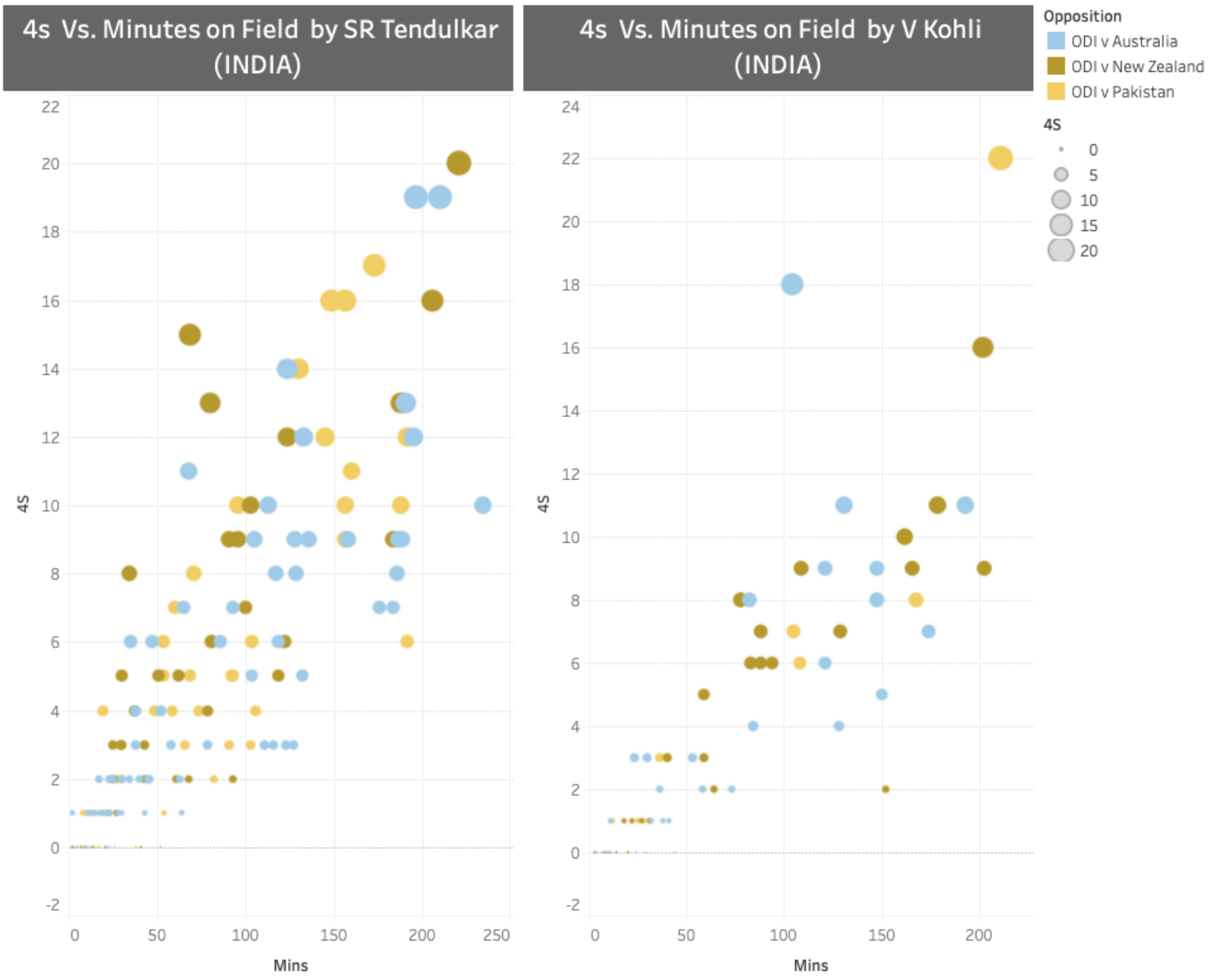


Fig 31. 4s scored vs minutes played by both players

⇒ Description

These visualizations show relationship between number of 6s scored and minutes spent of field when playing against Australia, Pakistan, New Zealand. The size of the circle represents number of 4s, and the color is used to distinguish between the opponents

⇒ Findings

- The number of 4s scored by Tendulkar is much higher than Kohli when playing against these opponents.
- In case of Tendulkar large number of 4s are concentrated between 120 and 200 minutes on field whereas in case of Kohli the number of 4s scored increases with minutes on field but remains almost the same from 100 minutes with a few outliers.

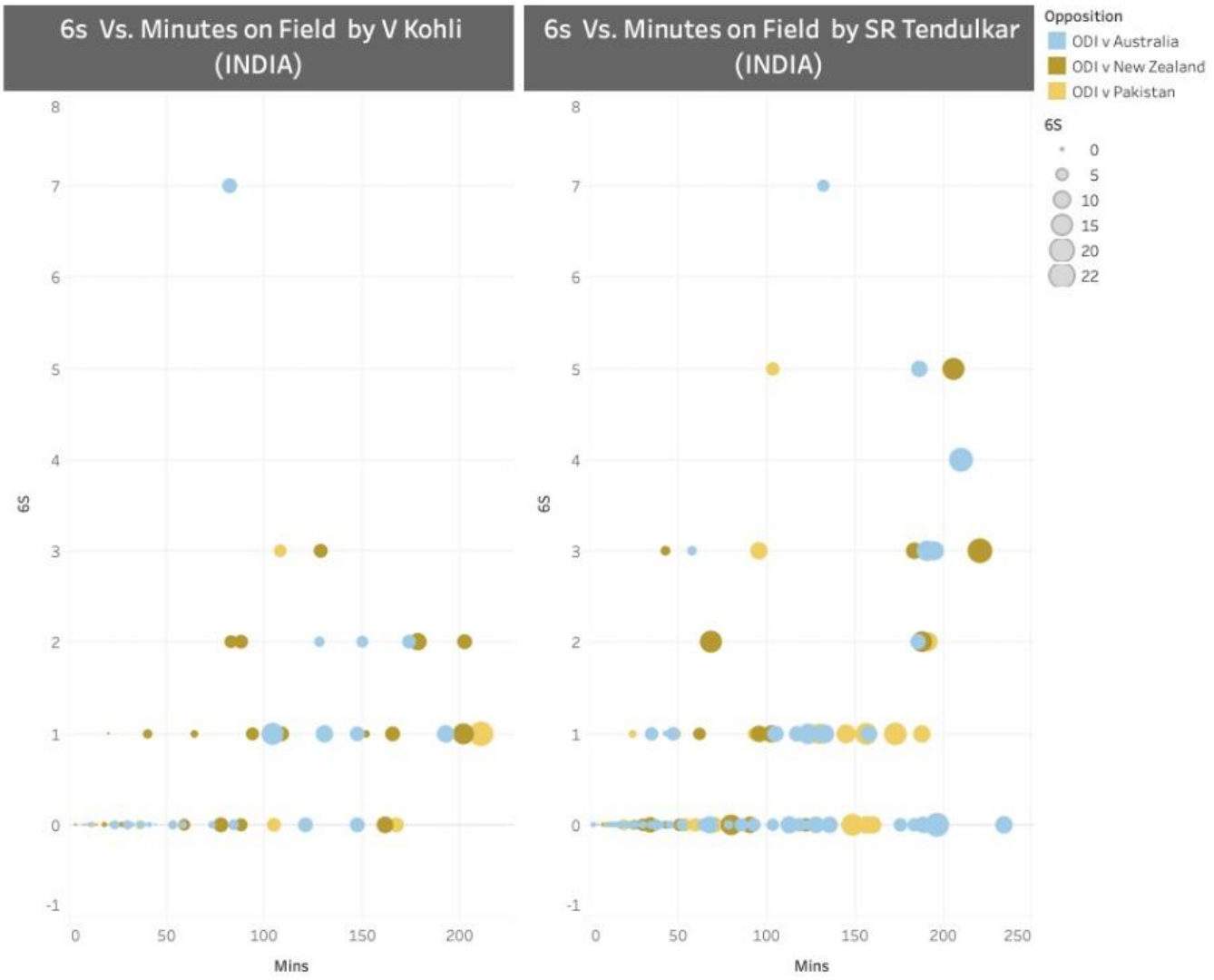


Fig 32. 6s scored vs minutes played by both players

⇒ Description

These visualizations show relationship between number of 6s scored and minutes spent of field when playing against Australia, Pakistan, New Zealand. The size of the circle represents number of 4s, and the color is used to distinguish between the opponents.

⇒ Findings

- The number of 6s scored by Tendulkar is much higher than Kohli when playing against these opponents.
- Tendulkar tends to score higher number of 6s starting from 100 minutes on field which increases as minutes on field increase whereas in case of Kohli large number of 6s are scored when number of minutes on field crosses 100 and from 100 minutes on field the number of 6s scored remains nearly the same for every match.

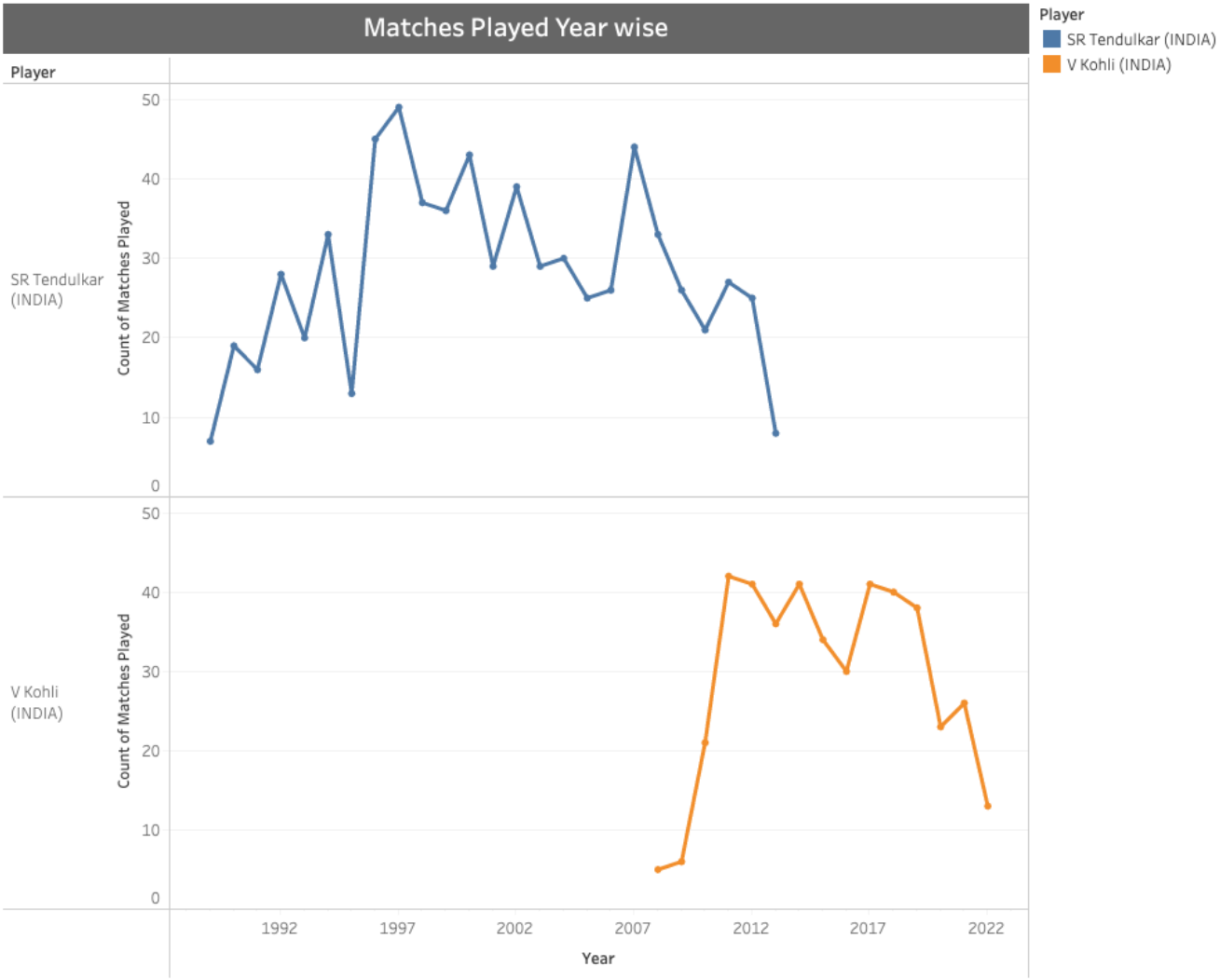


Fig 32. Matches Played yearwise by both players

⇒ Description

The visualizations show number of matches played by Sachin Tendulkar and Virat Kohli each year during their career.

⇒ Findings

- In 2011 and 2014, Virat played the maximum number of international matches while he played the least in 2022 and International Career started from 2008.
- In 1997 Sachin played the maximum number of international matches. Overall, the number of international matches played by sachin is higher than Kohli.

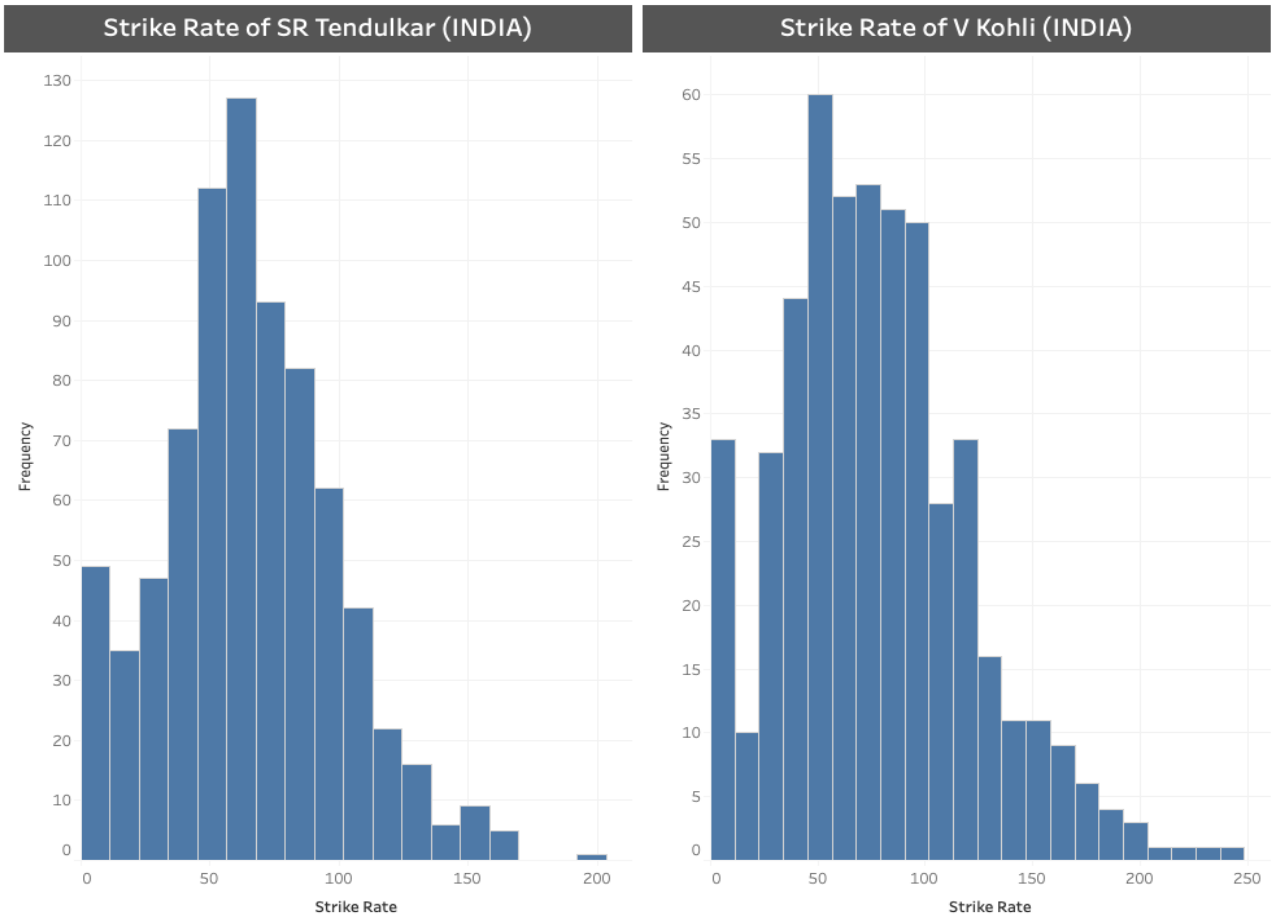


Fig 32. Match Strike rate both players

⇒ Description

Before we try to understand this metric let us first define it, Batting strike rate is a measure of how quickly a batter achieves the primary goal of batting, namely scoring runs, measured in runs per 100 balls; higher is better

The above visualizations show number of matches played by Sachin Tendulkar and Virat Kohli each year during their career.

⇒ Findings

- In 2011 and 2014, Virat played the maximum number of international matches while he played the least in 2022 and International Career started from 2008.
- In 1997 Sachin played the maximum number of international matches. Overall, the number of international matches played by sachin is higher than Kohli.

⇒ **Overall Summary for player vs player analysis**

Since his debut in 1989 till retirement, Sachin Tendulkar has played 463 ODI matches moreover he has managed to score **18,426 runs, 96 half-centuries, and 49 centuries,** He has also managed to score **1 double century in his ODI career.** Furthermore, the Master Blaster has **smashed 195 sixes and has an average of 44.83 in ODIs**. His highest score till date remains at **200 runs in ODI.**

On the other hand, current Indian skipper Virat Kohli made his **ODI debut in 2008**. Since then, he has played **245 matches and scored a massive 11,792 runs**. Unlike his senior, he hasn't scored any double century yet. However, he has **57 half-centuries and 43 centuries** to his name. Furthermore, he has hit **121 sixes** and has a **high score of 183**. Moreover, his ODI average is **59.86.**

In conclusion, looking at the statistics of Virat Kohli vs Sachin Tendulkar in ODIs, Sachin has set the bar high. However, Sachin has played many more games than Virat so it is yet to be seen if Virat can break the Master Blasters records. Therefore, as of now, Sachin leads according to ODI statistics of Sachin vs Virat.

Discussion of Findings

Based on the findings that we obtained in the previous sections here are some insights we were able to gather –

1) **The effect of home factor** (Section 3.1) –
    a. In ODIs home advantage has played a crucial factor in gameplay for Australia (70% win rate), Bangladesh, New Zealand.
    b. In T20s the away factor was surprisingly more favorable for India, Netherlands, New Zealand, South Africa, Sri Lanka and Zimbabwe.
    c. In test matches India, New Zealand, South Africa and Sri Lanka had distinctively evident home advantage.
    d. In terms of win margin in ODIs a general trend we observed is Teams have scored higher margins (>100 runs) with higher frequency (>3) against teams like Canada, Netherlands, Scotland, UAE.

2) **The rise in popularity of the sport of cricket** (Section 3.2) –
    a. We see a steady rise in the number of matches thus implying a rise in the popularity of cricket as a sport across the world in the past 2 decades.
    b. There is an exception in the year 2020 however, this is due to the COVID-19 pandemic which drastically reduced the number of games being played across the world

3) **Overall scoring trends in IPL** (Section 3.3) –
    a. We noted an overall increase in scoring trends in the last 5 years as opposed to the early years of the IPL, this is probably due to an improvement in the ability of players to score runs.

4) **Team level scoring trends in IPL** (Section 3.4) –
    a. We observed that that teams that score more runs, concede less runs on an average are usually the more successful ones.
    b. Scoring higher runs and conceding less runs is not the only way to win eg. in the year 2018 the team Chennai Super Kings won but had the highest runs conceded, to combat this they also scored highest runs that year.

5) **Is the toss a deciding factor in the team's performance in IPL ?** (Section 3.5) –
    a. On average, winning a toss seems to have no significant effect on a teams win/loss.
    b. We found that most teams opted to field first in the last 5 years, the trend was to bat first in the early years of IPL

c. We also found fielding first is favorable in order to win the game as most teams have a better win percentage by fielding first, this explains the reason for the above shift in toss decision.

6) **The effect of Boundaries on a team's performance in IPL** (Section 3.6) –
   a. we see that the number of boundaries is almost consistent across the years with about 10% increase in the latter years indicating that players have become better run scorers.
   b. Overs 6-10 seem to have the least number of boundaries scored.

7) **The effect of Scoring rate on game outcome in IPL** (Section 3.7) –
   a. winning teams score runs that are after the 3rd Quartile and losing teams seem to be more concentrated around the median and quartiles below that.
   b. We also observed a spike in runs scored in the 3rd and 4th overs and overs 18 – 20.

8) **The effect of venue on scoring in IPL** (Section 3.8) –
   a. The M. Chinnaswamy stadium (home stadium of the team Royal Challengers Bangalore) is among the high scoring stadiums. This could possibly explain why despite being a good team Bangalore had a surprisingly high number of average runs conceded per game.

9) Player comparison (Section 3.9) -
   a. In terms of the number of matches played, Tendulkar has played more matches than Kohli
   b. For every extra ball faced by Virat we can expect the runs to increase on average by a much higher amount than Sachin.
   c. The runs scored by Tendulkar is higher even when number of balls faced is small as compared to Kohli
   d. Kohli is the fastest of the two. Virat Kohli makes more runs in less time compared to Sachin Tendulkar.

References

1. Veroutsos, Eleni. "The Most Popular Sports in the World." *WorldAtlas*, WorldAtlas, 20 Oct. 2022, https://www.worldatlas.com/articles/what-are-the-most-popular-sports-in-the-world.html.
2. *Big data analytics- the new player in ICC World Cup Cricket 2023* (no date) *ProjectPro*. Available at: https://www.projectpro.io/article/big-data-analytics-the-new-player-in-icc-world-cup-cricket-2015/89 (Accessed: December 4, 2022).
3. Schroer, Alyssa. "From Fantasy Football Predictions to Baseball's Statcast, Big Data in Sports Is a Real Game Changer." *Built In*, 2018, builtin.com/big-data/big-data-companies-sports.
4. Jain, Sahil. "Technologies That Changed Cricket over the Years." *Sports News*, Sportskeeda, 22 June 2018, https://www.sportskeeda.com/cricket/technologies-that-changed-cricket-over-the-years/6.
5. *Red Ball Data*, 26 Apr. 2019, https://redballdata.blog/page/10/
6. *Kreedon*, https://www.kreedon.com/analysis-of-growth-in-ipl-viewership-a-case-study/?amp.
7. Singhal, Shashank. "Data Visualization - IPL Data Set (Part 2)." *Medium*, Analytics Vidhya, 30 Mar. 2021, https://medium.com/analytics-vidhya/data-visualization-ipl-data-set-part-2-b9daa59e4dc8.
8. Sunil. "Virat Kohli vs Sachin Tendulkar Stats: Who Is Better Batsman? (Comparison)." *CricketConnected.com*, 22 May 2022, https://cricketconnected.com/virat-kohli-vs-sachin-tendulkar-stats/