Decoding Cricket's Evolution: A Time-Travel through 1996-2005 Matches

Abhishek Reddy Gorla
Amulya Ambati
Lakshmi Prasanna Poluru
Venkata Harshith Nikhil Samudrala
Sai Vamsi Challamala
Sneha Sasanapuri

December 19, 2023

Abstract

Embark on a captivating journey through the transformative period of 1996-2005 in the world of cricket with our analysis. This paper delves into the statistical analysis of matches during this pivotal decade, uncovering the strategic shifts, iconic rivalries, and emergence of cricketing legends. Beyond mere scores, we dissect the geographical and chronological dimensions, shedding light on home-ground advantages, toss decisions. This comprehensive narrative also reveals patterns and trends that shaped the strategic evolution of teams.

Contents

1 Introduction

In the dynamic realm of cricket, where strategy, talent, and national pride converge, this exploratory data analysis endeavors to unravel the multifaceted narratives woven into the statistics of cricket matches spanning the years 1996 to 2005. Through an assortment of visualizations, we navigate the vast landscape of cricketing data, peeling back the layers to reveal compelling insights and patterns. From the acknowledgment of standout individual performances, visualized in a word cloud of 'Player of the Match,' to victories showcased on a world map, this analysis seeks to encapsulate the essence of a pivotal era in cricket. The exploration extends further into the nuances of ODI (One Day International) cricket scores through insightful box plots, scrutinizing the variance and trends that shaped the game during this time frame. Additionally, we delve into strategic decisions, examining the frequency with which teams chose to bat upon winning the toss, providing a unique perspective on the intersection of choice and outcome. Furthermore, a series-wise breakdown of team performance and a granular focus on India's contributions across multiple series deepen our understanding of the intricate tapestry of cricketing history during these transformative years.

2 Formulating Our Investigative Framework

This paper researches the influence of venue dynamics on cricket match outcomes and prompts the investigation of whether playing at home holds a substantial impact on a team's likelihood of securing victory, with a consideration of potential exceptions. It also investigates to see if there is an increased likelihood of winning the match for teams that win the toss. Additionally, we aim to discern patterns surrounding teams that demonstrate exceptional prowess or face challenges when competing in away matches.

3 Data

3.1 Introduction to Data Source

The data for this analysis originates from ESPN Cricinfo, the world's leading cricket website and among the top five single-sport websites in the world. ESPN Cricinfo serves as a reputable platform for cricket-related information, providing a robust foundation for the exploration and insights derived from the data set.

3.1.1 ABC's of Cricket: A Game of Bat and Ball

Cricket is a bat-and-ball game showcasing the skills and strategies of two teams, each comprising eleven players. Played on a field with a 22-yard pitch and a wicket at each end, the game involves bowlers propelling the ball towards the wicket while batters aim to score runs by striking the ball and running between the wickets. This paragraph provides a concise yet comprehensive overview of the fundamental dynamics of the game.

3.1.2 Structure and Dynamics of One Day Internationals (ODIs)

An One Day International (ODI) is a distinctive form of limited-overs cricket characterized by a fixed number of overs (currently 50) and lasting up to 9 hours. This format, i.e., a limited-overs competition is recognized for its significance in international cricket, with major tournaments such as the Cricket World Cup adhering to the ODI structure.

3.2 Data Collection Methodology

Cricinfo receives live cricket data from multiple sources, including official data feeds from cricket boards, broadcasters, and on-site reporters. These data sources provide information about every ball bowled, runs scored, wickets taken, and other match-related details. And they also have their dedicated media zone at the live cricket match center where they sit and keep the scores updated just like how they update on the digital scoreboard at the stadiums.

3.3 Primary attributes for analysis

The data set under analysis comprises a comprehensive collection of cricket match records spanning the period from 1996 to 2005. Each entry in the data set is characterized by a set of key features, offering a rich source of information for cricket enthusiasts and analysts alike. The primary attributes include:

- Match Details: Unique identifiers (match_id) for each match, participating teams (team1 and team2), and relevant match dates (match_date)
- Score and Wickets: Records of the runs scored (score_team1 and score_team2) and wickets taken (wickets_team1 and wickets_team2) by each team
- Match Conditions: Details such as venue (venue), and the chosen option after winning the toss (toss_decision)
- Series Information: Information about the cricket series (series) under consideration, allowing for series-wise analysis of team performances
- Outcome and Margin: The match winner (winner), the margin of victory (margin), and the corresponding margin type (margin_type)
- Player Recognition: The players honored with "Player of the Match" title (player_of_match), along with their affiliated team (player_of_match_team).
- Toss Details: The team winning the toss (toss) and the number of balls remaining at the conclusion of the match (ball_remaining)
- Geographical Context: The geographical context is provided through attributes such as the ground (ground), the city of the ground (ground_city), and the country in which the match took place (ground_country) indicating home or away for teams

3.4 Potential sources of bias

The cricket dataset encompassing ODIs played between 1995 and 2006 may be influenced by several important biases. Matches played during this period may be subject to era-specific norms, potentially limiting the generalizability of conclusions to more recent cricketing trends. Additionally, the dataset may exhibit biases favoring historically stronger teams, specific tournaments, and certain venues, influencing analyses related to team strength, tournament dynamics, and home-ground advantage. The over-representation of well-known players and matches with specific outcomes could introduce biases in individual player performance and match result analyses. Series-wise assessments may be biased by disproportionate representation from specific bilateral series, while variations in data reporting across sources may impact the accuracy and completeness of analyses. Furthermore, historical rule changes and inadequate consideration of weather conditions pose additional biases that should be carefully addressed for a nuanced and accurate interpretation of the cricket dataset.

4 Methods

4.1 Data Cleaning

4.1.1 Dealing with Missing Values

We have handled missing values in certain columns like 'ball_remaining', 'margin', 'margin_type', 'series' with values [0, 0, 'default', 'NatWest'] so that this approach replaces NAs to ensure uniformity in the dataset. According to internet resources based on venue and date of match, we figured out that matches whose series name is missing are part of 'Natwest Series'. For all other remaining columns, we have proceeded ahead with default values. By employing this method, we successfully minimized information loss and preserved the true characteristics of the data.

4.1.2 Parsing Date Column

The date columns in the dataset were subjected to parsing and standardization to ensure consistency and facilitate meaningful temporal analysis. This process involved converting the date information from its original format into a standardized date format that is easily interpretable and compatible with analytical tools.

4.2 Exploratory Data Analysis(EDA)

Our goal is to unearth meaningful stories embedded in the data and set the stage for more targeted analyses. We commenced our EDA by presenting descriptive statistics that offer a summary overview of the dataset's central tendencies, standard deviation, and distribution. These statistics serve as a preliminary glimpse into the key numerical attributes, guiding us in identifying trends and potential outliers. By embarking on this EDA journey, we aim not only to uncover the inherent characteristics of the dataset but also to lay the groundwork for informed decision-making in subsequent phases of our data exploration.

Using the data insights acquired through the analysis of different features, the creation of various info graphics, and the incorporation of our existing knowledge along with general cricket-related inquiries, we formulated several hypotheses. Subsequently, we conducted hypothesis testing to assess the statistical significance of these hypotheses. The following section delves into a detailed discussion of the tested hypotheses.

Upon observing the visualization 9, it is evident that India's batting average varies against different countries. Consequently, we intend to conduct a hypothesis test to further explore and analyze these variations.

1. Does India's batting average exhibit a notable difference when playing against Pakistan compared to its batting average against South Africa?

   - Null Hypothesis: The batting average of India against Pakistan equals the batting average of India against South Africa.
   - Alternate Hypothesis: The batting average of India against Pakistan is not equal to the batting average of India against South Africa.

   We employed a two-sample mean test methodology for conducting the aforementioned analysis. Focusing on India's batting average, we specifically included matches where India batted first. This approach was chosen because, in cases where India bats second, the team only needs to match the opponent's score, which doesn't accurately reflect the true batting average potential.

   After filtering the matches as suggested before, the sample size became small and we ended up with less than 30 matches for both India vs Pakistan and India vs South Africa. Hence we couldn't proceed with z-test. We tried to check for the normality of the data to be able to use t-test.

   - Plot a histogram of the batting scores to see if it follows a normal distribution.
   - Check if the data has a symmetric nature and is not too spread out similar to normal distribution using mean - min $\approx$ max - mean $\approx$ k * sd

   The statistic to be computed is $\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma}{m}+\frac{\sigma}{n}}}$

   However, we used the inbuilt function present in python 'statsmodels.stats.weightstats.ttest_ind' to perform the test.

2. Is there an increased likelihood of winning the match for teams that win the toss?

   - Null Hypothesis: The ratio of matches won and the toss won to the total number of matches where the toss was won is the same as the ratio of matches won and the toss lost to the total number of matches where the toss was lost.

- Alternate Hypothesis: The ratio of matches won and the toss won to the total number of matches where the toss was won is larger than the ratio of matches won and the toss lost to the total number of matches where the toss was lost.

Since we are looking at whether a data point satisfies a condition or not P(X=1) = p and P(X=0)=1-p and it follows a bernoulli distribution. Since our sample size is large we can approximate the mean to follow normal distribution.

Hence, we used a two-sample proportion test using the z-statistic methodology for carrying out the above analysis.

The statistic to be computed is $\frac{\hat{p}_{won\_match\_won\_toss} - \hat{p}_{won\_match\_lost\_toss}}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n}+\frac{1}{m})}}$

However, we used the inbuilt function present in python 'statsmodels.api.stats.proportions_ztest' to perform the test.

3. Is there a variance in a country's likelihood of winning when playing at a neutral ground (a location not belonging to either of the participating countries) compared to playing at the home ground of the opponent's country?

   - Null Hypothesis: The proportion of matches won by a country on neutral ground is equal to the proportion of matches won on the opponent's home ground.
   - Alternate Hypothesis: The proportion of matches won by a country on neutral ground is greater than the proportion of matches won on the opponent's home ground.

   We used a two-sample proportion test methodology for carrying out the above analysis.

   The statistic to be computed is $\frac{\hat{p}_{won\_match\_won\_toss} - \hat{p}_{won\_match\_lost\_toss}}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n}+\frac{1}{m})}}$

   However, we used the inbuilt function present in python 'statsmodels.api.stats.proportions_ztest' to perform the test.

   In conducting this analysis, we have chosen four countries based on the number of matches available in the dataset and their overall performance. The selected countries include India, England, Zimbabwe, and New Zealand.

5 Results

5.1 Infographics

5.1.1 Player of the Match Showcase

This word cloud provides a captivating visual summary of the standout players who have earned this prestigious accolade across various cricket matches. Each player's name is represented in the word cloud, with the size of the text corresponding to the frequency of their recognition as the player of the Match. This engaging visualization not only showcases the diversity of players who have excelled in different matches but also highlights any recurring names that have consistently made a significant impact on the game. The players who won this title most of the times were **Sourav Ganguly**, **Sachin Tendulkar**, **Sanath Jayasuriya**, **Jacques Kallis** and **Ricky Ponting**.

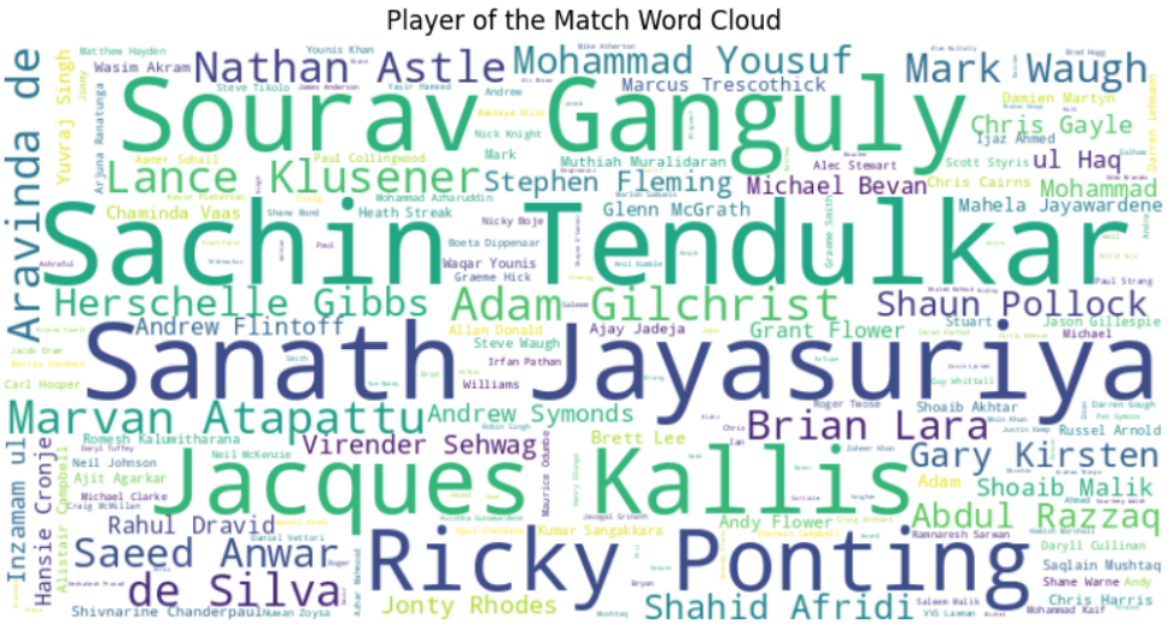


Figure 1: Cricket stars - Player of the Match from 1996 to 2005

5.1.2 Players of the Match: A Visual Celebration

In this bar graph, we have shown the exceptional performances of the top 5 cricket players who have received the prestigious "Player of the Match" award. Each horizontal bar corresponds to a player, showcasing their award numbers with bold clarity. The accompanying country flags add a patriotic touch, visually connecting each player to their cricketing nation. The graph also infuses a sense of national pride as we recognize and applaud the cricketing prowess of these remarkable players. The nations to which these players belong to were India, Sri Lanka, India, Zimbabwe and Australia in the order of appearance of bars.

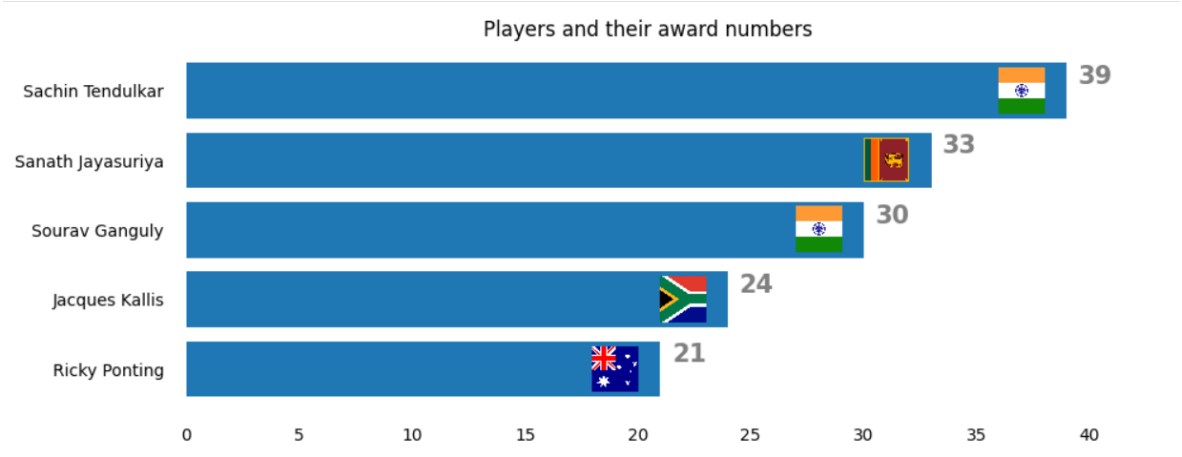


Figure 2: Top 5 Players of the Match: Award Frequencies and Country Flags

5.1.3 Global Cricket Dominance: Choropleth Map of ODI Match Wins (1996-2005)

In our exploration of One Day International (ODI) cricket matches spanning the years 1996 to 2005, we turn our attention to the global stage to understand the distribution of victories among participating

countries. To visually capture the dominance of cricketing nations during this period, we present a compelling choropleth map, allowing for an intuitive and comprehensive view of the global distribution of ODI match wins. The color intensity on the map corresponds to the number of matches won by each country. Darker shades indicate higher victory counts. This visual representation aligns with our overarching goal of gaining a comprehensive understanding of ODI cricket dynamics, setting the stage for deeper analyses and insights in subsequent sections of our report.

- Notably, **South Africa** and **Australia** appeared with lighter color, indicating an exceptional performance in terms of match wins.

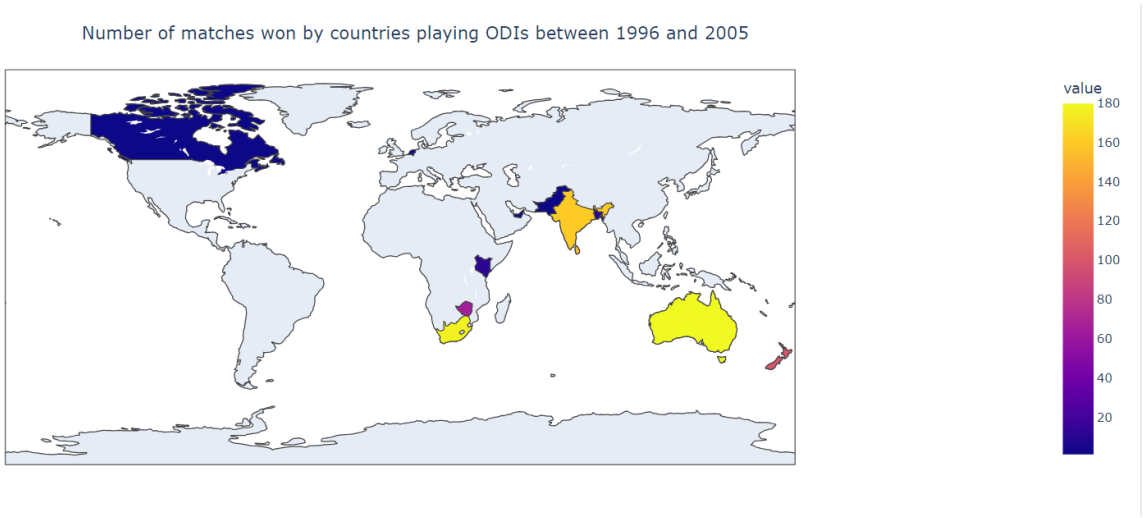


Figure 3: Global Distribution of ODI Match Wins (1996-2005)

5.1.4 Exploring ODI Cricket Scores Across Countries: A Box Plot Analysis

In our quest to unravel the dynamics of ODI cricket, we turned our attention to the scoring patterns of various cricketing nations. This box plot provides a holistic view of each country's batting performance. The box itself encapsulates the IQR, signifying the spread of scores within which the central 50% of data lies. This helps in illustrating the interquartile range (IQR) for each country, offering insights into the spread and variability of scores. The line within each box represents the median score, offering a robust measure of the typical performance for each country. A wider box suggests greater variability in a team's performance. Countries with smaller IQRs suggest a more consistent performance in terms of runs scored.

- **Australia** demonstrated a robust median score, and a relatively small box compared to 'West Indies' and 'Namibia' indicating consistent performance across matches.
- **New Zealand** displayed a notable outlier of 397 runs, indicating an exceptional batting performance in 'Videocon Triangular Series' match that significantly deviates from the norm.
- **Canada** encountered the least favorable performance with a score lower than 50 runs against Sri Lanka is an example of where the ball took the centre stage unlike the bat.

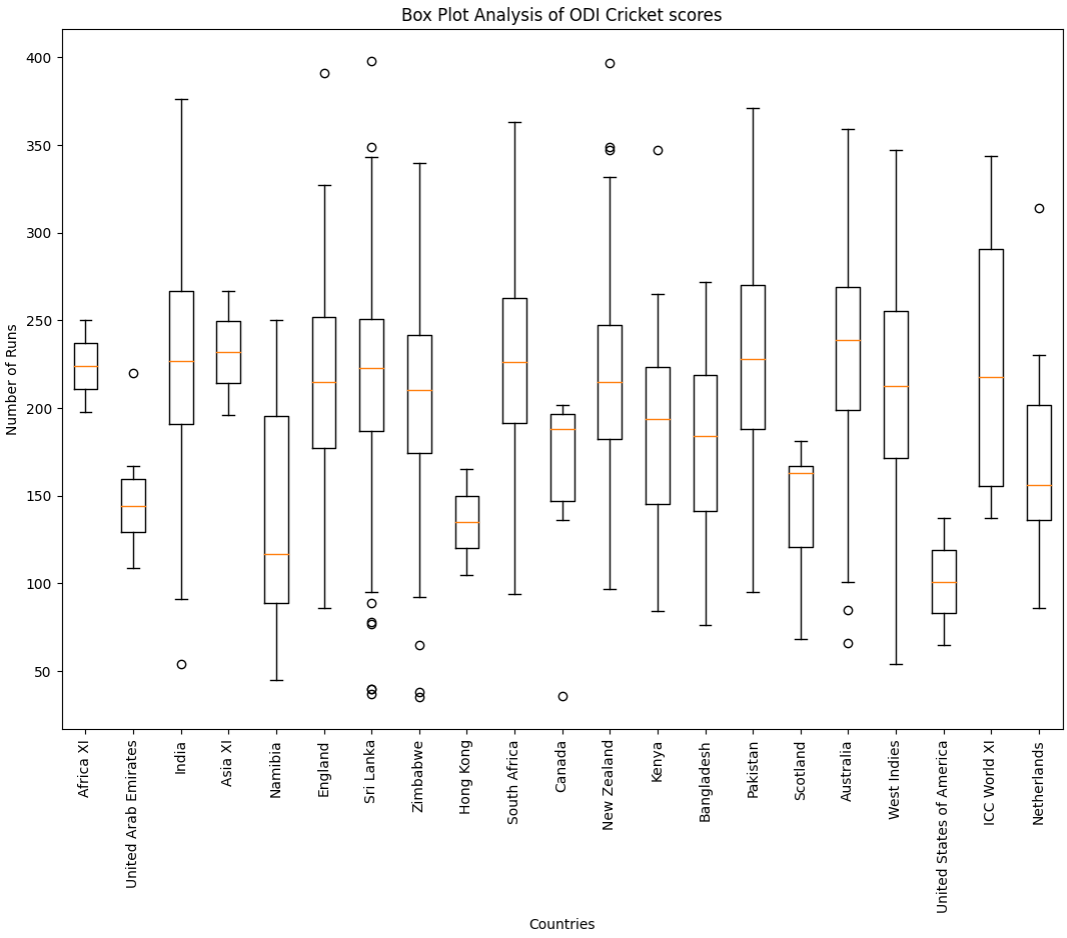


Figure 4: Box Plot Analysis of ODI Cricket Scores Across Countries

5.1.5 Toss Strategy: A Comparative Analysis of Batting Choices

By computing the percentage of times each country chooses to bat, we gain valuable insights into the prevailing strategies employed by cricketing nations. The percentage values provide a quantitative measure of the inclination towards batting as the preferred choice. The chart vividly highlights variations in the batting choices of different countries. Some nations consistently favor batting, while others exhibit a more balanced or varied approach.

- Countries like **Pakistan, Australia, Bangladesh** and **Kenya** tended to opt 'bat first' for almost 70% of the times they won toss.
- **West Indies** and **Sri Lanka** displayed an even distribution in their decision-making, opting to 'bat first' approximately 50% of the times they won the toss.

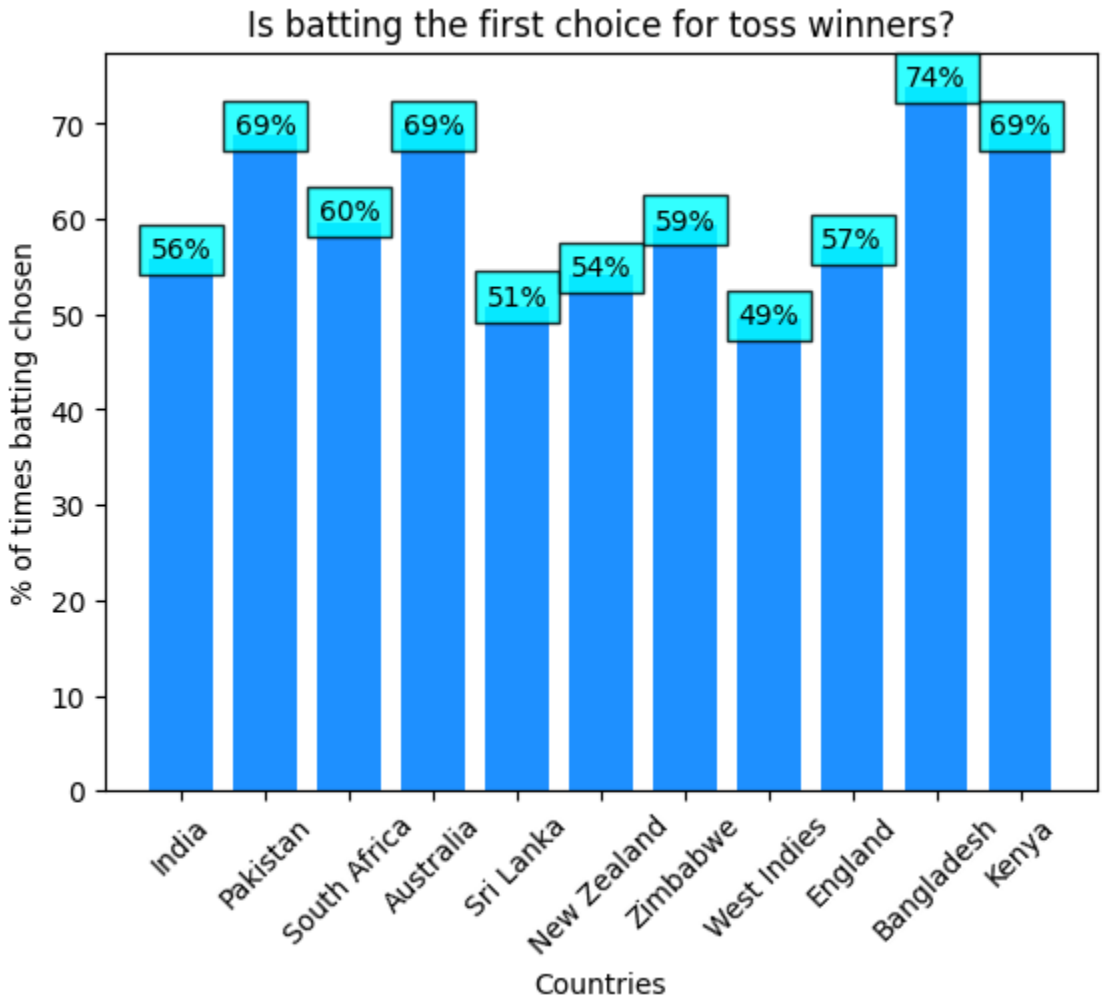


Figure 5: Batting Preferences among Toss Winners: A Country-wise Analysis

5.1.6 Series-wise Analysis of Team Performance

The chart visually presents the distribution of wins for countries that have played at least 15 matches across different cricket series. Each horizontal bar is stacked, representing the cumulative wins for a team in a particular series. The length of each segment indicates the number of victories achieved in that specific series.

- **Australia, Pakistan and South Africa** emerged as dominant forces, consistently securing victories in various series.
- **Australia and India** exhibited remarkable leadership in ICC World Cup, showcasing their prowess in adapting to varying match conditions.
- Also, **Australia** secured a significant number of victories in the VB series and Carlton & United series, both of which took place on their home turf.

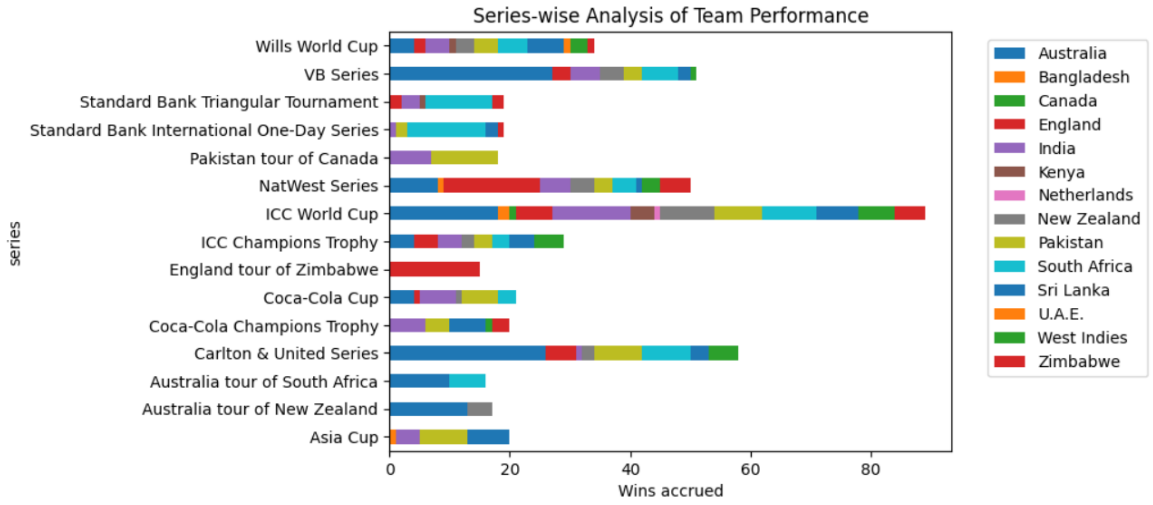


Figure 6: Batting Preferences among Toss Winners: A Country-wise Analysis

5.1.7 Most formidable opponent on the field

A team either wins by scoring more runs (if it wins by batting first) or by wickets (if it wins after batting second). The team that bats second is called the chasing team. Win margin is measured in two forms:

- By Runs: Variation in the run total when the team batting first secures a victory.
- By Wickets: The variance in the wicket count when the chasing team emerges victorious. This is attributed to the fact that the match concludes as soon as the chasing team successfully pursues the target set by the team batting first.

In terms of runs: Matches with the lowest win margins(by runs) suggest a closely contested battle between the two teams. This indicates that the outcome of the game was uncertain until the very end, making it more exciting for both players and spectators. The presented visualizations highlight the matches with thrilling conclusions, where the outcome was determined by a minimal run difference, emphasizing the competitiveness of these matches. This visual reveals the win margin of batted-first countries in ODI matches spanning from 1996 to 2005.

- **Zimbabwe** had proven to be a formidable adversary, featuring in the top ten most fiercely contested matches on four occasions.
- Countries involved in these matches demonstrate a capacity to engage in closely fought battles, showcasing their resilience and ability to perform under pressure.

Figure 7: Lowest Winning margins: By Runs

In terms of wickets: This visualization unveils the narrative of countries batting second in successful chases. In cricket, if a team wins by a margin of 10 wickets, it means that the team successfully chased the target without losing any wickets. This implies that both openers (the first two batsmen in the batting order) remained not out and achieved the required target without the team losing a single wicket. It is considered a comprehensive and dominant victory, indicating a strong batting performance by the winning team.

- **Bangladesh** had lost thrice among top 10 successful chase matches. This shows either poor batsmanship or opponent's excellent bowling spirit.

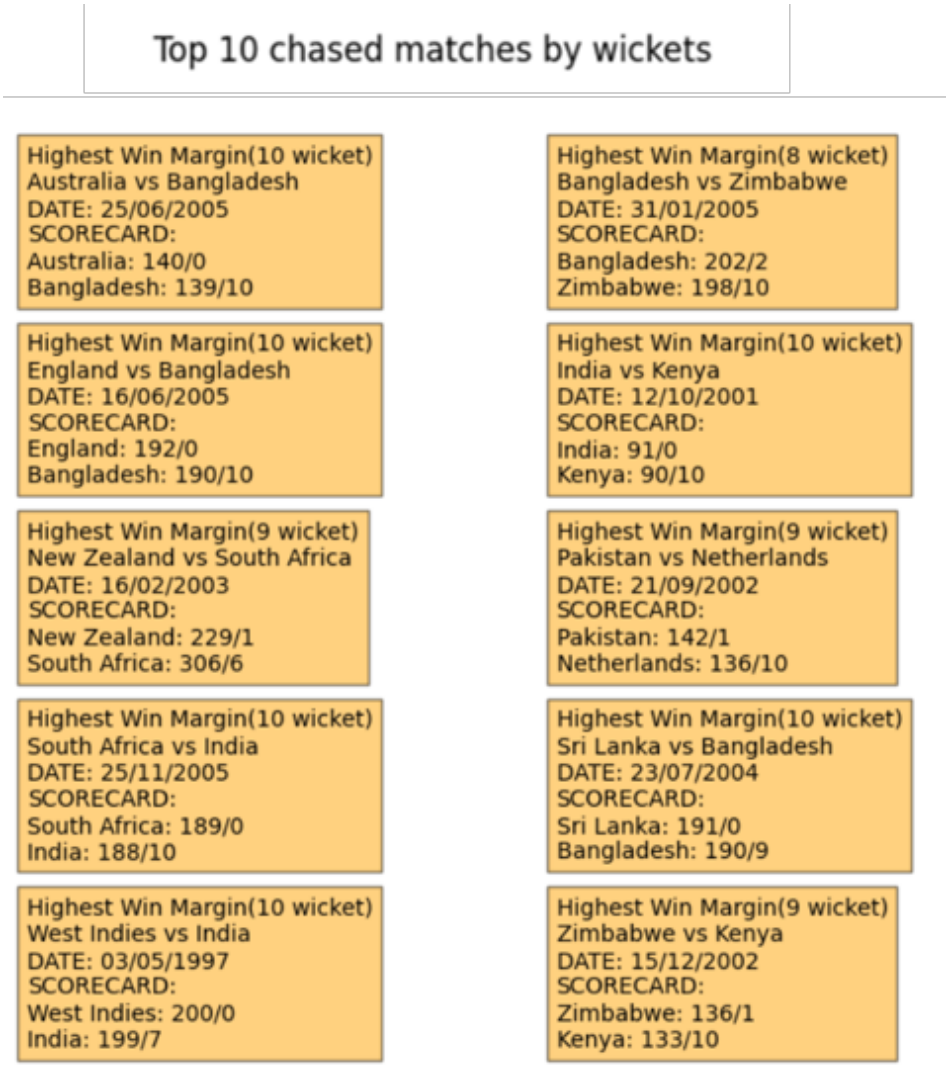


Figure 8: Highest Winning margins: By Wickets

5.1.8 India's Batting Performance

The Solid Gauge chart below offers a concise and visually impactful representation of India's batting performance against multiple countries. Matches in which India had played more than 5 times against a country and batted first were only considered for analysis. Each segment of the gauge corresponds to the average runs scored by the Indian cricket team in a specific series during this decade. Each arc in the chart represents a series, and the length of the arc corresponds to the performance score for that series.

- India exhibited its highest batting averages when facing **Zimbabwe, Sri Lanka**, and **Australia**. An exploration into the concealed influence of **home** matches, where the advantage of pitch familiarity comes into play, was undertaken to scrutinize the widely accepted notion that teams generally experience a notable advantage when playing on their home turf. The table presented below showcases India's performance against these three nations in both home and away matches. A significant difference was observed w.r.t home and away matches for a specific country. For instance, average runs were 279 against Australia for home matches but 225 for away matches.

Opponent country	Avg runs	No of matches	Category
Zimbabwe	283	10	home
Zimbabwe	216	3	away
Sri Lanka	276	5	home
Sri Lanka	242	5	away
Australia	279	6	home
Australia	225	5	away

Table 1: India's Batting Performance: Home vs. Away matches against Zimbabwe, Sri Lanka & Australia

.

- India demonstrated comparatively lower batting averages against **England** and **New Zealand** when contrasted with their performance in other countries. Upon closer examination of the table below, compelling evidence emerges, suggesting that the impact of home matches plays a significant role in shaping India's performance.

Opponent country	Avg runs	No of matches	Category
England	281	1	home
England	200	5	away
New Zealand	309	4	home
New Zealand	143	5	away

Table 2: India's Batting Performance: Home vs. Away matches Against England & New Zealand

.

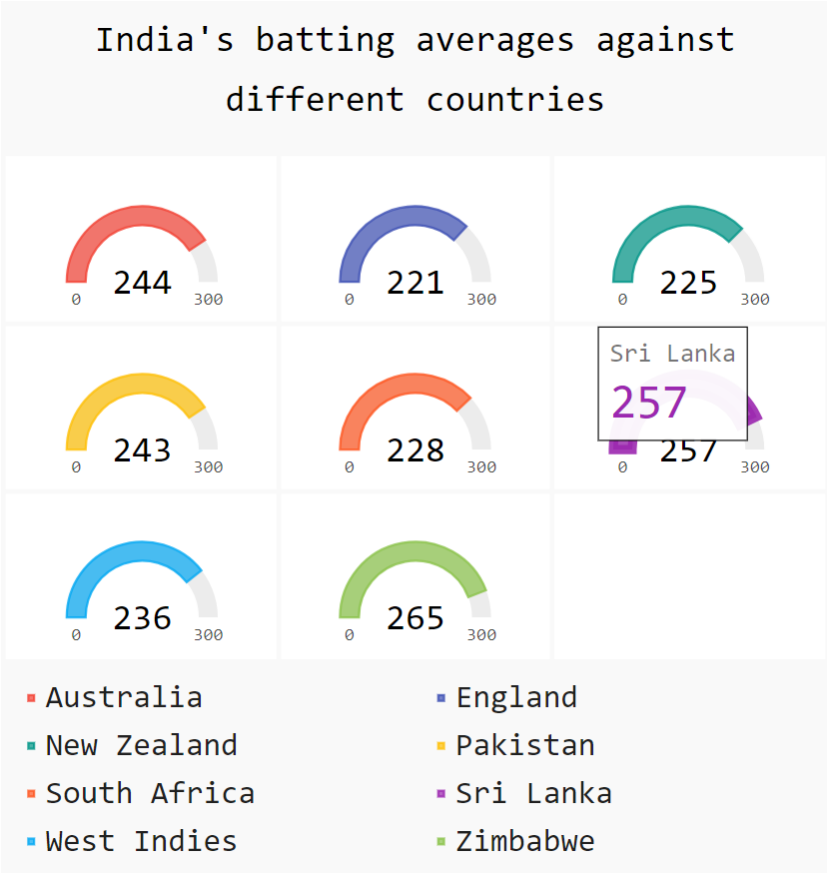


Figure 9: Average runs scored by India when Batted First in more than 5 matches

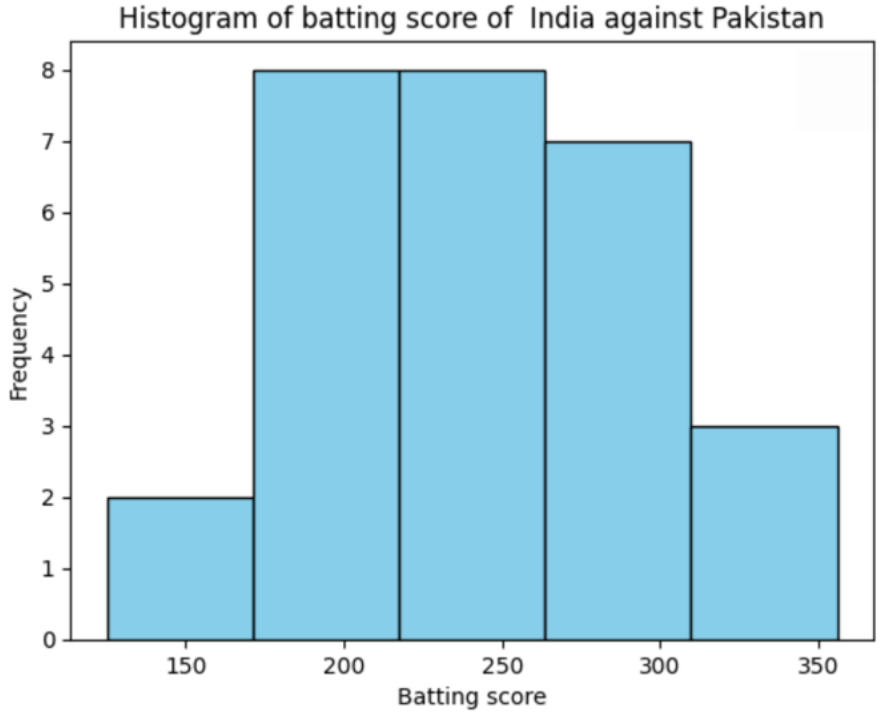
5.2 Statistical Tests

1. Does India's batting average exhibit a notable difference when playing against Pakistan compared to its batting average against South Africa?

   $H_0 : \mu_{Ind\_Pak} = \mu_{Ind\_SA}$

   $H_A : \mu_{Ind\_Pak} \neq \mu_{Ind\_SA}$

   By following the approach discussed in section 4.2 we did a normality check by looking at the symmetric nature and spread of the data.



Opponent country	(Mean-Min)/sd	(Max- Mean)/sd
Pakistan	2.17	2.06
South Africa	1.48	1.81

Table 3: Spread of the data for normality check

   From the above metrics and visual representation, we concluded that the data in the sample is normally spread and t-test can be applied to it. Here the data assumption is that scores of Ind vs Pak and Ind vs SA are approximately normally distributed.

   We set the significance level to 0.05 and performed the two-sample mean hypothesis test.

   The obtained p-value was 0.29, hence we failed to reject the null hypothesis.

   Hence we couldn't prove that India's batting average against Pakistan is significantly different from India's batting average against South Africa.

2. Is there an increased likelihood of winning the match for teams that win the toss?

   $H_0 : \hat{p}_{won\_match\_won\_toss} = \hat{p}_{won\_match\_lost\_toss}$

   $H_A : \hat{p}_{won\_match\_won\_toss} > \hat{p}_{won\_match\_lost\_toss}$

   We set the significance level to 0.05 and performed the two-sample proportion hypothesis test.

   The obtained p-value was 0.30, hence we failed to reject the null hypothesis.

   Hence we couldn't prove that toss winners have a higher chance of winning the match.

3. Is there a variance in a country's likelihood of winning when playing at a neutral ground (a location not belonging to either of the participating countries) compared to playing at the home ground of the opponent's country?

   $H_0 : \hat{p}_{won\_at\_neutral} = \hat{p}_{won\_at\_away}$

   $H_A : \hat{p}_{won\_at\_neutral} > \hat{p}_{won\_at\_away}$

   We set the significance level to 0.05 and performed the two-sample proportion hypothesis test.

   p-value for the hypothesis on 4 different countries are as follows

Opponent country	p-value	Decision
India	0.015	Reject null hypothesis
England	0.019	Reject null hypothesis
Zimbabwe	0.36	Fail to reject null hypothesis
New Zealand	0.12	Fail to reject null hypothesis

Table 4: Result of hypothesis testing across various countries

   We can see that for India and England, as the p-value is less than the significance level, the proportion of wins at the neutral ground is greater than the opponent's home ground, whereas it's not the case for Zimbabwe and New Zealand.

   Hence we could say that the hypothesis is true for a few countries but is not a generic one.

References

1. Github page for data
2. ESPN Data source