Unveiling Insights through Data-driven Exploration of Cricket World Cup 2023

Team Members - Anshuman Mohanty (USC ID- 4257570790), Kaustubh Sharma (USC ID-1765749035)

Introduction

The recently concluded ICC Men's 50 Over Cricket World Cup 2023 holds untapped data that can provide strategic insights. Our project aims to evaluate the performance of all the teams across the tournament by analyzing various aspects such as individual batting and bowling performances, venue-specific trends, the impact of toss decisions on match outcomes, team powerplay tactics, and first and second-innings performances. We want to understand how these factors contributed to some teams' success and others' underperformance in the tournament. To accomplish this, we collected data by scraping web data from reliable cricket news websites like ESPNCricinfo and Cricbuzz using the BeautifulSoup and Requests Python modules. We then applied appropriate data cleaning methods to transform the raw data to help us with our subsequent analysis steps. Finally, we transformed our analysis results to visualizations to communicate our findings effectively.

1. Data Collection

We systematically gathered data from reputable sources such as CricBuzz, and ESPNCricinfo. Employing the BeautifulSoup and Requests Python modules facilitated efficient web scraping, enabling the extraction of pertinent information from the respective websites. The resulting dataset was meticulously organized and stored in csv files, laying the groundwork for a thorough visualization and analysis of match statistics. Our project comprises of data from the group stages of the tournament (45 matches in total) and best individual batting and bowling performances of the entire tournament.

Our collection process starts with collecting the data on team standings, facts related to each match like match date, team scores, team names, match venue, links to match scorecard and winner of each match. As this data was in table format located inside div classes, we extracted the required data from HTML elements inside <table> tags, like <tr>, <thead>, <tbody>, <th> and <td> within a particular <div> class. The links to each match scorecard was taken from the <a> tag.

Upon iterating every match link, we had collected powerplay-wise (3 powerplays) runs and wickets distribution for each team in each match. Furthermore, each player's batting performance was meticulously recorded, encompassing details like player name, runs scored, balls faced, total 4s and 6s, and strike rate in the order of their batting position.

Parallel to the batting scorecard, our exploration delved into the bowling statistics for every group-stage match. The collected data details include the total number of overs bowled, maiden overs, runs conceded, economy rate, dot balls, 4s & 6s conceded, wides, and no balls for each bowler in every team.

Finally, we analysed the best individual batting and bowling performances throughout the tournament such as highest runs scored, best batting strike rate and average, highest wickets, best economy rate and bowling average.

The Batting and bowling scorecards along with the data on best individual performances were structured as tables inside div elements and hence, were scraped in a similar way as match facts mentioned earlier.

Note: The team standings (points_table.csv) was collected from Cricbuzz and the remaining data was collected from ESPNCricinfo.

2 Data Cleaning

Following the scraping of the data, we undertook several vital steps throughout the data cleaning process to refine our dataset for optimal analysis and visualization. Initially, given the extensive information scraped from the web and stored in multiple csv files, we carefully selected the most important files upon which our analysis and visualization would be based on, thus discarding the unimportant files. Further, we had dropped a few columns which were not of much use for our analysis like match date, victory margin etc. During the scraping process, few on records were populated with non-ASCII characters which were removed using regular expressions.

Following the removal of ASCII characters, several records had data attached with value within parentheses (e.g.: - Player Name (Country Names), Country Name (Qualification Status), Team Score (Number of Overs) etc.). On similar lines, there were some records were two parts of data separated with a ‘/’ (for e.g.: - 200/3, here 200 is the runs scored and 3 is the wickets lost). So, the value inside the parenthesis and value after the ‘/’ was extracted with help of regular expressions from the existing column and was placed under a new column to enhance our dataset’s clarity.

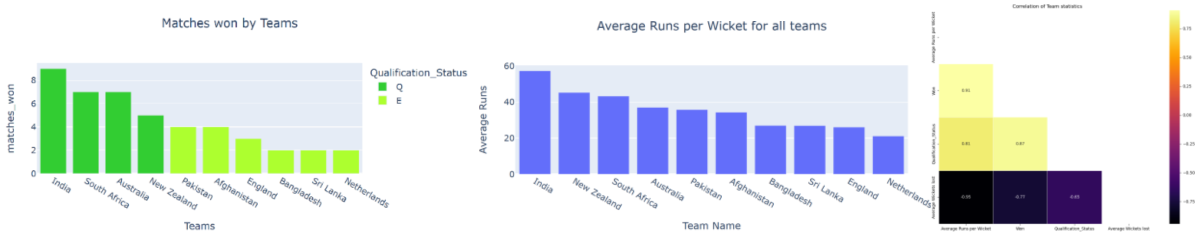
A central file (match_summary.csv) was constructed by merging multiple csv files containing specific match details like teams involved, team scores, toss data, venue, Man of the Match, powerplay wise wickets and runs distribution, winner of the match etc.

Upon performing a null value check, we observed missing data in batting scorecard file (not all 11 batsmen get a chance to bat every game) and the match summary file (teams had missing data for powerplay 3 either since teams lost all their wickets or successfully chased the target within the first two powerplay phases). So, the missing batting data was filled with zeroes and missing powerplay 3 data was filled with ‘NA’.

After scraping the bowling data, bowling statistics such as economy rate, dot balls etc. observed to be in the float data type, which was converted to integer type to ensure consistency.

3 Analysis and Visualizations

This section highlights the different types of statistical analysis and inferences from the processed csv files. Using the extracted data, we have computed average run scored per wickets lost, run rate, economy rates, strike rate etc, phases-by-phase average runs scored and wickets lost.



(a) (b) (c)

Fig 1: Bar plot containing (a) matches won by teams (b) average runs scored per wicket. (c) Heatmap showing correlation of the general team statistics

From the figure 1(a), we can see that India, South Africa, Australia, and New Zealand are the best performing teams as they have qualified for the knockout stages. A similar trend is observed with 1(b) as these top-performing teams have the best average runs scored per wicket. This fact is further confirmed by the heatmap in 1(c) as we can see a high positive correlation of winning teams with runs scored per wickets lost and a high negative correlation with the average wickets lost per match which, in turn lead to the qualification of these teams to the knockouts.

Further, to break down our analysis we have performed the analysis to find out the best performing teams while Batting 1st (Bowling 2nd) and vice versa.

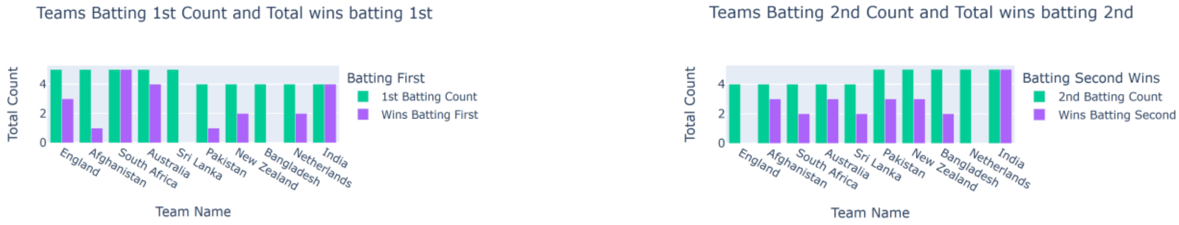
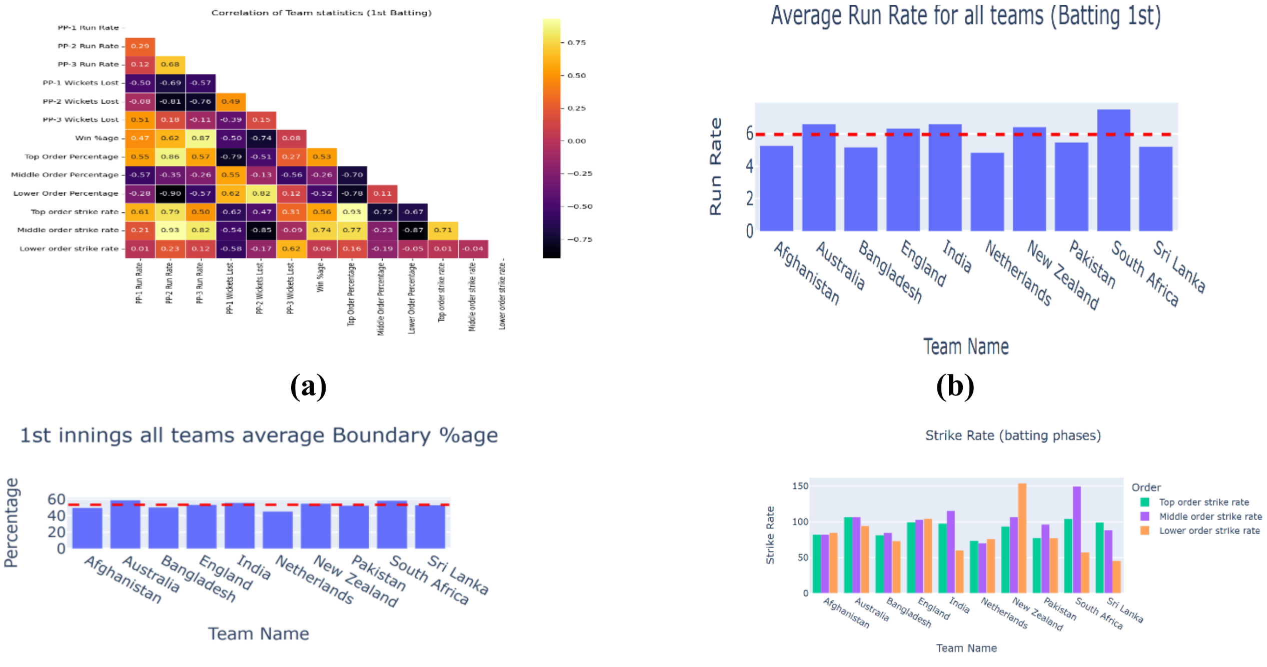


Fig 2: Bar plot containing Match played and won by teams (a) Batting 1st (b) Batting 2nd

From the figure 2 (a), we can see that India and South Africa have won all their matches while batting first, followed by Australia, England and New Zealand being the part of top five performing teams (batting first/bowling second). Similarly, while chasing (batting second/bowling first), it is observed that India has won all their matches. Afghanistan and Australia have lost one match each followed by Pakistan and New Zealand who have lost two while batting second are the part of top 5 chasing teams.



(a) (b)

(c) (d)

Fig 3: Batting 1st (a) Correlation Statistics (b) Average Run rate (c) Boundary percentage (d) Strike Rate

From the figure 3 (b), (c) and (d), it can be observed that the top 5 teams batting first have had the highest average run rate, boundary percentage and strike rate. The same is confirmed by the heatmap from the figure 3(a). Further, it was observed from the heatmap that teams who have won more matches (batting 1st) are observed to have a higher Middle Order Strike Rate and PP-3 Run Rate. There is a significant positive correlation between Top Order Contribution and PP-2 rate indicating that teams whose top order has had higher contributions have achieved a higher rate of scoring in PP-2. Also, teams who have lost more wickets in PP-2 are observed to have a lower scoring rate in PP-2 and PP-3, go on to lose the match.

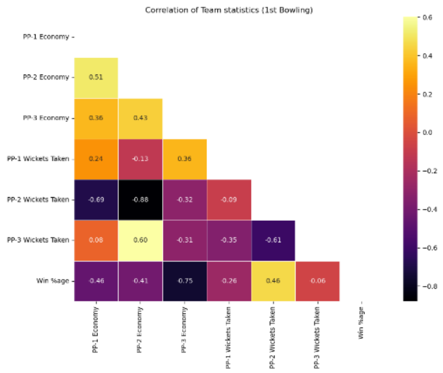
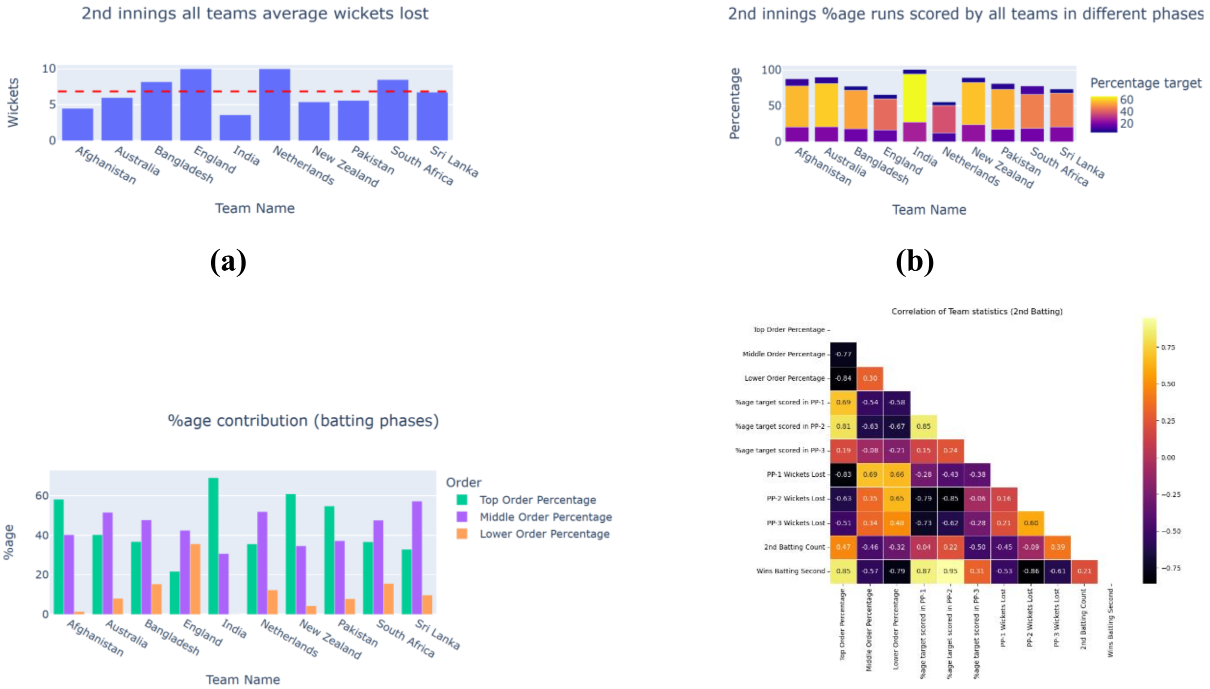


Fig 4: Bowling 1st Correlation Statistics

Similarly, from the teams who have won bowling first, it is observed that teams have conceded the less runs in PP-3 and have taken more wickets in PP-2 which is inferred from the heatmap.

Similarly, moving on the second innings analysis



(a)

(b)

Fig 4: Batting 2nd (a) Average Wickets lost (b) percentage target runs achieved in powerplay phases (c) Batting order percentage contribution (d) Correlation Statistics

From the figures 4 (a), (b), (c) and (d), it can be clearly seen that best chasing teams have lost the least amount wickets per match, scored highest percentage of the target in PP-2, and had the highest top order contributions. It is further confirmed by the heatmap in 4 (d). Additionally, it was observed that teams whose top order scores more runs is observed to have lose fewer wickets during PP-1.

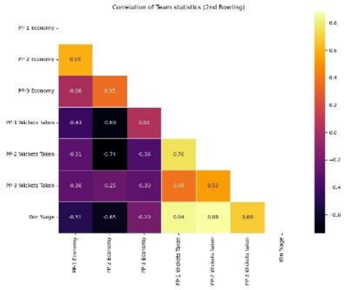


Fig 5: Bowling 2nd Correlation Statistics

From the figure 5, it can be concluded that teams bowling second have taken a high number of wickets in all three powerplays. Teams who take these wickets manage to control the scoring rate of the opposition and go on to successfully defend the score and win the match.

Note: This notebook contains the most relevant visualizations, all other visualizations can be found in the file 'Inferences for Visualization.ipynb'.

Conclusion and Future Scope:

This project on Data analysis on Cricket World Cup 2023 provides a detailed insight on why some of the teams have performed well and others haven't. Many conclusions were drawn from our study. Firstly, it shows that teams batting first and bowling second are successful when the batsman score runs at a quicker rate and bowlers manage to pick wickets throughout the bowling innings. Similarly, it could be observed that teams batting second and bowling first are successful when they have high contributions from their top order and take wickets in PP-2 during their bowling innings which reduces the run flow. If we extend our project timeline, we will have the opportunity to delve deeper into the exciting final stages of the ICC Cricket World Cup, including the crucial semi-final and final matches. These moments are where the tension peaks and exploring them could add a layer of insight into how teams perform under heightened pressure, providing a more comprehensive analysis. Also, we would also like to refine our statistical models, perhaps even introducing advanced machine learning algorithms. This step aims to unlock more nuanced insights from the dataset, enabling us to predict match outcomes and uncover intricate player performance trends. When it comes to visualizing our findings, the extra time will allow for the refinement of visual representation techniques. We would explore interactive dashboards using tools such as PowerBI or Tableau, aiming to convey the intricate details of cricketing statistics in an engaging and accessible manner.