

## **Cricket Match and Player Analysis**

Shaunak Sarkar

202184354

MSc Data Analytics

Supervisor: Dr Lindsey Corson

26/08/2022

## **Declaration**

Except where explicitly stated, all the work in this dissertation – including any appendices – is my own and was carried out by me during my MSc course. It has not been submitted for assessment in any other context.

Signed:

Name: Shaunak Sarkar

Date: 26/08/2022

## **Ethics**

This dissertation has been prepared using cricket match data of the tournament Rachael Hayhoe Flint Trophy primarily from the website <https://cricsheet.org/> and other online portals. These websites are legal. The data is free and publicly available and contains necessary information of every match played. None of the data is sensitive and invasive to privacy of any individual. The data has been procured with full knowledge and permission of my supervisor. The project also does not involve any live participant.

This project is ethically compliant.

## **Project Context**

## List of Contents

Declaration.....	i
Ethics.....	ii
<b>Project Context</b>	1 - 4
1 Background.....	1
2 Project Purpose.....	2
3 Project Plan.....	2
4 References.....	4
<b>Client Report</b>	1 - 49

## 1. Background

Cricket is a popular sport played using a bat and a ball between two teams comprising eleven players on each side. The game is played on an oval shaped ground with a rectangular 'pitch' of 22 yards in the middle. A set of three sticks called 'wickets' are placed on each end of the pitch. Two horizontal items called 'bails' are placed on top of the wickets. The two playing teams take turns to bat and bowl, the order of which is decided on the basis of the outcome of a coin toss. Each such turn is referred to as an 'innings'. For the batting team, two players will be on the field at a time on either side of the 'pitch'. One batter will be the striker who will face the ball being bowled by the bowler from the other end. The other batter will be on the non-striker end which is the same side of the bowler. The batter has to score 'runs' by hitting the ball with the bat and then running to the opposite end of the pitch. The batter at the non-striker end also must do the same immediately. When the batters successfully reach their positions on the other ends of the pitch, one run is scored. Likewise, if they can do the same twice then two runs are scored. A batter may choose not to run and thereby not score in that ball. A batter can also score runs by hitting the ball successfully towards the stadium boundary. Four runs can be scored if the ball reaches the boundary line after hitting the ground. Six runs are scored when the ball directly hits the boundary or beyond without bouncing anywhere within the ground. A batter is 'out' or has to stop playing when her 'wicket' is taken. A 'wicket' refers to any situation in general when a batter is out. A batter is 'bowled out' if the incoming ball hits the wicket sticks and displaces the bails. A batter can also 'caught out' if the ball after being hit is caught by an opponent player before reaching the ground. A 'run out' happens when the bowl hits the wicket sticks on any end of the pitch before a batter reaches that end of the pitch while running. Other ways of getting out are LBW (leg before wicket) or field obstruction or retired hurt due to injury. When a batter is out, the next player in that team comes to bat. Each team is allowed maximum ten wickets during an innings. The purpose of the batting team is to score runs and ensure that fewer wickets fall while doing so. For the bowling team, all players are on the field during the innings. One of them bowls to the opponent batter while the rest are put on fielding positions to restrict the movement of the ball after being hit by the batter. The bowling team has to bowl a certain number of overs based on the match format. Six successive balls constitute an over and have to be completed by the same bowler. An innings concludes when all overs have been completed. The bowling team aims to take as many wickets as possible and restrict the number of runs scored by the opponent. There are certain deliveries like wide ball or no ball which are not allowed by the bowling team and in such cases the batting team is awarded 'extra' runs and extra balls may have to be bowled in that over. The outcome of a cricket match is decided during the second innings. A team wins either when it scores runs that exceed the total runs scored by the opponent in the previous innings or when it restricts the opponent by taking wickets and keeps their runs below the required. (Williams, Alston and Longmore, 2021)

Regarded as England's national summer sport, cricket is widely played in Commonwealth nations of India, Pakistan, Sri Lanka, Bangladesh, Australia, New Zealand, South Africa and West Indies. There is also rising interest in the sport in several other countries including Afghanistan, Nepal, UAE, Oman, Scotland, Ireland, Canada and the USA. Various multi-nation tournaments like the World Cup or Champions Trophy as well as two-nation or three-nation series are conducted throughout the year. Often team selections for these competitions are influenced by the performance of players in domestic tournaments within the country. For example, there are five domestic tournaments conducted by the England and Wales Cricket Board inclusive of both men and women's cricket.

## **2. Project Purpose**

Big data has taken over the world of sports and cricket is no exception. Every cricket match produces ball-by-ball data: for each batsman we can record the number of balls faced, the number of runs scored, the strike rate, the number of fours and sixes, etc. For each bowler we can record the number of overs bowled, the number of wickets that the bowler has taken, the number of runs given, etc. Collectively, all this data has the potential to create vast opportunities to make meaningful insights.

This project involves using publicly available data to explore match and player data from professional cricket tournaments to determine how key statistics relate to match results. Analysis of these statistics may be used to determine the factors most strongly associated with winning matches.

The tournament of concern in this project is the Rachael Heyhoe Flint Trophy, a women's cricket domestic tournament of England that is conducted in the 50 over format. Available data from the years 2020 and 2021 will be used to study and statistically analysed to draw inferences on significance of factors contributing to winning cricket matches.

## **3. Project Plan**

The project will be completed over a period of 12 weeks starting from 7<sup>th</sup> June 2022. An expected weekly schedule has been prepared during the first week that is given on the following page. It takes into account the possible time period for execution and completion of multiple required tasks inclusive of technical analysis and preparation of the report. The schedule plan also takes into account the expected dates of meetings with supervisor for discussing the progress of work, any impediments and feedback sessions. The volume of work assigned on a weekly basis has been estimated realistically by taking into consideration any unexpected situation or delay. The distribution of tasks has been done in a balanced manner in order to complete them successfully by 31<sup>st</sup> August 2022.

Week												
1	2	3	4	5	6	7	8	9	10	11	12	13
07-06-2022	13-06-2022	20-06-2022	27-06-2022	04-07-2022	11-07-2022	18-07-2022	25-07-2022	01-08-2022	08-08-2022	15-08-2022	22-08-2022	29-08-2022
First meeting with supervisor												
• Background research of topic												
• Literature review												
	<ul style="list-style-type: none"> <li>Study of Data set</li> <li>Initial stages of analysis: data preparation, data cleaning</li> <li>Commencement of exploratory data analysis</li> </ul>		Second meeting with supervisor	<ul style="list-style-type: none"> <li>Action on feedback/ recommendation</li> <li>Continue exploratory analysis</li> </ul>								
					Third meeting with supervisor	<ul style="list-style-type: none"> <li>Action on feedback/recommendation</li> <li>Continue exploratory analysis</li> <li>Gathering insights from analysis</li> </ul>						
							Fourth meeting with supervisor					
							<ul style="list-style-type: none"> <li>Action on feedback/recommendation</li> <li>Continue analysis and reporting of findings</li> <li>Reflect on analysis procedure and improve /redo/add required changes</li> <li>Prepare report draft</li> </ul>					
										Submit report draft to supervisor for review		
										<ul style="list-style-type: none"> <li>Action on feedback/recommendation</li> </ul>		
											Final meeting with supervisor	
											<ul style="list-style-type: none"> <li>Prepare final report with necessary changes</li> </ul>	
												Submit report by 31/08/2022



#### **4. References**

Williams, M.K., Alston, R. and Longmore, A. (2021). cricket. In: *Encyclopedia Britannica*. [online]  
Available at: <https://www.britannica.com/sports/cricket-sport> [Accessed 11 Jun. 2022].

## **Client Report**

## List of Contents

1 Executive Summary.....	1
2 Acknowledgement.....	2
3 Introduction.....	3
3.1 Women's cricket and its growth.....	3
3.2 Rachael Heyhoe Flint Trophy.....	4
4 Literature Review.....	4
5 Process Overview.....	6
6 Exploratory Analysis Observations.....	7
6.1 Toss and Match Winners.....	7
6.2 Winning Teams and Winning Margins.....	8
6.3 First Innings Runs.....	9
6.4 Individual Batter Runs.....	11
6.5 Batting Strike Rate.....	12
6.6 Batting Average.....	13
6.7 Batting Partnerships.....	14
6.7.1 Partnership Networks of different teams and runs scored.....	15
6.7.2 Partnership Networks and Runs of Southern Vipers and Sunrisers per match.....	19
6.8 Wickets and Wicket-takers.....	22
6.9 Bowling Economy Rate.....	24
6.10 Bowling Strike Rate and Bowling Average.....	24
7 Identifying Key Players.....	26
8 Scoring Players and Teams.....	27
9 Predictive Modelling.....	30
9.1 Predicting Match Winners.....	30
9.1.1 Using Team Batting and Bowling Scores.....	30

9.1.2 Using Toss Outcome.....	32
9.2 Predicting Second Innings Runs.....	33
9.2.1 Using Team Batting and Bowling Scores.....	33
9.2.2 Using Information of First 10 overs .....	36
9.3 Discussion .....	37
10 Conclusion .....	38
11 References.....	39
12 Appendices.....	43
Appendix 12.1: Top 20 runs scoring batters in an innings.....	43
Appendix 12.2: Top 10 overall runs scoring batters.....	44
Appendix 12.3: Top 20 runs scoring batting partnerships in an innings.....	44
Appendix 12.4: Top 10 overall runs scoring batting partnerships.....	45
Appendix 12.5: Top bowlers with most wickets (4 or above) in an innings.....	46
Appendix 12.6: Top bowlers with most wickets (10 or above) overall.....	46
Appendix 12.7: Python code for Linear SVC modelling.....	47
Appendix 12.8: Python code for Linear Regression modelling.....	48
Appendix 12.9: Python code for Ridge Regression modelling.....	49

## List of Figures

Figure 1a: Toss Decision.....	7
Figure 1b: Match Outcome of Toss Winners.....	7
Figure 1c: Toss Decision of Toss and Match Winners .....	7
Figure 1d: First Innings state of Match Winners .....	7
Figure 2: Number of Wins Teamwise.....	8
Figure 3: Winning margins by runs and wickets.....	9
Figure 4: First Innings Runs Distribution.....	10
Figure 5: Highest and Lowest First Innings Runs.....	10
Figure 6: Individual Batter Runs in an Innings and Overall.....	11
Figure 7a: Distribution of Batting Strike Rates in an Innings.....	13
Figure 7b: Distribution of Mean Batting Strike Rates.....	13
Figure 8: Distribution of Batting Average.....	13
Figure 9: Partnership Runs in an innings and overall.....	14
Figure 10: Partnership Overall Network and Net Runs of Southern Vipers.....	15
Figure 11: Partnership Overall Network and Net Runs of Western Storm.....	16
Figure 12: Partnership Overall Network and Net Runs of Lightning.....	16
Figure 13: Partnership Overall Network and Net Runs of Northern Diamonds.....	17
Figure 14: Partnership Overall Network and Net Runs of Sunrisers.....	17
Figure 15: Partnership Overall Network and Net Runs of Central Sparks.....	18
Figure 16: Partnership Overall Network and Net Runs of Thunder .....	18
Figure 17: Partnership Overall Network and Net Runs of South East Stars.....	19
Figure 18: Partnership Network of each match of Southern Vipers .....	20
Figure 19: Partnership Network of each match of Sunrisers.....	21
Figure 20: Distribution of Partnership Runs in an Innings of Southern Vipers and Sunrisers.....	22
Figure 21: Wicket Types.....	23
Figure 22: Wickets Taken by Bowlers in an Innings.....	23

Figure 23: Highest Wicket taking Bowlers.....	24
Figure 24: Economy Rate in an Innings and Overall.....	24
Figure 25: Bowling Strike Rate and Average in an Innings and Overall.....	25
Figure 26: Distribution of Player Batting and Bowling Scores.....	28
Figure 27: Distribution of Team Batting and Bowling Scores.....	29
Figure 28: One-hot encoding.....	30
Figure 29: SVM Classification (Saini, 2021).....	30
Figure 30: Confusion Matrices of Linear SVC models (predicting winner by scores).....	31
Figure 31: Linear SVC models: 2022 Winner Predictions based on Scores.....	32
Figure 32: Confusion Matrices of Linear SVC models (predicting winner by toss outcome).....	33
Figure 33: Linear SVC models: 2022 Winner Predictions based on Toss Outcome.....	33
Figure 34: Linear Regression (Abhigyan, 2020).....	34
Figure 35: Linear Regression and Ridge Regression model Predictions based on Scores.....	35
Figure 36: Linear Regression and Ridge Regression model Predictions based on First 10 overs.....	36

## List of Tables

Table 1: Linear SVC model test accuracies (predicting winner by scores).....	31
Table 2: Linear SVC model predictions of 2022 winners (predicting winner by scores).....	32
Table 3: Linear SVC model test accuracies (predicting winner by toss outcome).....	32
Table 4: Linear SVC model predictions of 2022 winners (predicting winner by toss outcome).....	33
Table 5: Linear Regression and Ridge Regression model test accuracies (predicting by scores).....	34
Table 6: Linear Regression and Ridge Regression Second Innings Runs Predictions(based on Scores)....	35
Table 7: Linear Regression and Ridge Regression model test accuracies (predicting by first 10 overs).....	36
Table 8: Linear Regression and Ridge Regression Second Innings Runs Predictions (based on first 10 overs).....	37
Table 9: Top 20 runs scoring batters in an innings in 2020.....	43
Table 10: Top 20 runs scoring batters in an innings in 2021.....	43
Table 11: Top 10 overall runs scoring batters in 2020.....	44
Table 12: Top 10 overall runs scoring batters in 2021.....	44
Table 13: Top 20 runs scoring batting partnerships in an innings in 2020.....	44
Table 14: Top 20 runs scoring batting partnerships in an innings in 2021.....	45
Table 15: Top 10 overall runs scoring batting partnerships in 2020.....	45
Table 16: Top 10 overall runs scoring batting partnerships in 2021.....	45
Table 17: Top bowlers with most wickets (4 or above) in an innings in 2020.....	46
Table 18: Top bowlers with most wickets (4 or above) in an innings in 2021.....	46
Table 19: Top bowlers with most wickets (10 or above) overall in 2020.....	46
Table 20: Top bowlers with most wickets (10 or above) overall in 2021.....	47

## **1. Executive Summary**

Cricket is one of the most popular sports in the world (Shvili, 2020). It associates itself with multiple numeric statistics which in turn provides ample scope to explore different analytical approaches. Women's domestic cricket tournaments have gained immense popularity and Rachael Heyhoe Flint Trophy is one such tournament, recently started in 2020.

This project puts its focus on match data of the 2020 and 2021 seasons and analyses various aspects of batting and bowling. The impact of toss, runs scored by teams, runs scored by individual players, batting partnerships, strike rates, wickets, economy rates, etc have been studied and thoroughly observed for insights. Top performing players of every team were identified and their impact for team performances were looked into.

Using results of exploratory analysis, attempts were made using machine learning algorithms to find ways to predict the winning team of a match and the total runs of a team batting second. Both classification and regression approaches have been explored and their results were validated with the actual results of the 16 matches that have been already played till 23rd July for the 2022 edition of the tournament.



## **2. Acknowledgement**

I would like to thank my supervisor Dr Lindsey Corson for helping me understand the project context and purpose as well as providing essential feedback during the meetings and being always available for resolving potential roadblocks during every stage of the project work.

### **3. Introduction**

#### **3.1 Women's cricket and its growth**

Around the time of the eighteenth century, cricket was predominantly played by men as a sport although women occasionally participated in the game with a more recreational approach. After the game everyone enjoyed themselves by drinking and dancing. The rules of the game were not standardised at that time and formulated by local bodies. The first officially recorded women's cricket match was played in England in 1745. Women's involvement in the game were often influenced by the prevalent social norms like class, family restrictions and limitations of accepted feminine behaviour by male-dominated society. It further declined in the nineteenth century when cricket was being brought under more structured rules and regulations and was being associated with the national identity of England. During the first world war and the subsequent period thereafter, women had begun to participate in more male-centric jobs and their views were given more considerations in decision making processes. Women got better access to education which in turn gave more access of variety of sports that included cricket. However, women's cricket continued to stay under the lens of conservative cultural norms and restrictions as well as male dominance (Velija, 2015). The England Women Cricket Association (EWCA) was formed in 1926 and after navigating through years of financial and administrative issues, it merged with the men's organisation, the England and Wales Cricket Board (ECB) in 1998. Likewise in 2005, the International Women's Cricket Council (IWCC) merged with the international men's cricket governing body the International Cricket Council (ICC) (Velija, Ratna and Flintoff, 2012).

Slowly yet steadily there has been rise in popularity in women's cricket. This has resulted in associated increase in viewership as well as sponsorship that in turn facilitates the expenditure to improve infrastructure and scout for talent. In 2013, BBC reported about a study conducted by the England and Wales Cricket Board (ECB) which found that over 60000 women were engaged in playing cricket in the UK (BBC, 2013). The women's T20 World Cup tournament in 2020 reportedly was viewed on digital platforms 1.1 billion times and had a total spectator count of 136,549. The final match was itself watched by 86,174 people in the stadium (ICC Media Release, 2022). Women's cricket in the 20-over format has also been included in the Commonwealth Games 2022 to be held in Birmingham (England and Wales Cricket Board, 2019). Rising enthusiasm is also evident for domestic cricket tournaments. The women's edition of The Hundred, a professional 100 ball tournament started by the ECB in 2021 recorded high number of views both in person and through broadcast. The opening game was the watched by a record number of audience of 1.95 million which made history for being the most watched match in women's cricket (Mathews, 2021). Currently along with The Hundred the ECB organises two other regional women cricket competitions - the Rachael Heyhoe Flint Trophy in the 50 overs format and the Charlotte Edwards Cup in the 20 overs format.

### **3.2 Rachael Heyhoe Flint Trophy**

The Rachael Heyhoe Flint Trophy is a 50 over domestic women's cricket competition involving eight regional English teams. The tournament is named after former English cricketer Rachael Heyhoe Flint who passed away in 2017. Her sporting career spanned from 1960 to 1982 of which she served as the English team's captain from 1966 to 1978. She led her team to win the 1973 Women's Cricket World Cup a tournament which she also helped in organising. She relentlessly campaigned for the promotion and acceptance of women's cricket tackling administrative and financial hurdles. She was the first woman cricketer to be admitted to the International Cricket Council's hall of fame in 2010 (Nicholson, 2021).

In the 2020 season, the eight teams were divided into two groups of four teams each. Teams in a group played against each other twice and earned points per win. The final was played between top team of each group. The first edition of the Rachael Heyhoe Flint Trophy was won by the team Southern Vipers with runners-up the Northern Diamonds. The 2021 season did not have any groups and all teams played against each other once. The top team with most points from wins directly proceeded towards the final. The second and third teams had to qualify for the final through a playoff match between them. The 2021 season was also won by the Southern Vipers with runners-up the Northern Diamonds (Wikipedia, 2022a). The current 2022 season will be played from 2nd July till 25th September 2022. The format will be same as that of the 2021 season.

## **4. Literature Review**

Like any sport, in cricket the main objective of both the playing teams is to win the match which primarily depends on the goodness of batting, bowling and fielding. Cricket is a game that produces a lot of numerical entities like score, runs, overs, balls bowled, bowling speed, net run rate, wickets taken, catches and extras. Other subjective factors surrounding a cricket match can be pitch condition, weather conditions like rain, dew or wind, wear and tear of the ball being used, familiarity of venue, crowd support, etc. Often sports experts, analysts, commentators and speculative fans have referred to such numerical and other qualitative aspects in order to predict the winner of a game. Over the numerous games that have been played over the years, there have been some results where the winning team could be predicted during the second innings itself and there have been many nail-biting finishes where none could decide the winner till the last ball had been bowled. The several data surrounding a match have also been useful for post-match analysis and to study the sport in general by interested statisticians and researchers.

Many previous works have looked into different statistical methods to analyse and predict cricket match result or to compare and formulate patterns of individual teams' and players' performances. According to Bandulasiri, Brown and Wickramasinghe (2016), other than the on-field performance with the bat and the ball, the outcome of a match is also influenced by certain decisions like picking the most suitable combination of eleven players for the match, accurately analysing the changing pitch conditions, planning the batting line up and fielding positions as well as selecting the right bowlers for the crucial overs. Bailey and Clarke (2006) identified a range of variables that could statistically help in predicting the runs and match result. Some of them were team or player specific such as match experience, past performances and current form while some were venue specific like home ground advantage, performance and experience at that specific venue. Performance against a specific opponent team was also considered. They applied multiple linear regression to find out the probable margin of victory and the first innings score for a team. Their method also looked into situations where the playing overs were reduced due to rain. Saikia (2020) calculated the overall strength of a team and quantified a team's current form using two parameters- BP (Batting Performance) and ACBR (Adjusted Combined Bowling Rate). Adhikari, Saraf and Parma (2017) formulated a Dominance factor for team management of domestic leagues to assess bowlers by utilizing two parameters- RBW (Runs conceded Between consecutive Wickets) and BBW (Balls bowled Between consecutive Wickets). Mukherjee (2013) employed network analysis techniques to study batting partnerships. Lewis (2005) proposed a potentially fairer Duckworth/Lewis methodology to evaluate players' performances that took into consideration the stages of matches when runs are scored and wickets taken or lost. Parag Shah and Mitesh Shah (2015) developed a logistic regression model incorporating variables like ICC points of playing teams, their previous match results, team batting first, home ground advantage, time of match (day/night), etc to find the most probable team to win the match. Veppur Sankaranarayanan (2014) with the help of both historical and current match data, combinedly used linear regression and nearest neighbour classification aided attribute bagging algorithm to predict the runs scored at the end of innings during various key points in a match and thereby the winner. Kampakis and Thomas (2015) used multiple team and player performance statistics and applied Naïve Bayes and Random Forest algorithms to predict game outcomes of twenty over English County matches. Naïve Bayes was also used by Wickramasinghe (2020a) to predict the winning team of fifty over one-day international matches. Along with Naïve Bayes, Dubey, Suri and Gupta (2021) explored other machine learning algorithms like KNN, logistic regression, SVM and random forest classifiers in order to predict the winner. Wickramasinghe (2020b) studied the classification of allrounder players into different categories by utilizing Naïve Bayes, KNN and random forest classifiers. Multiple machine learning algorithms were also used by Passi and Pandey (2018) to find out the possible number of runs to be scored or wickets to be taken by players of both sides. Generally, in studies involving multiple machine learning estimators, ensemble methods-based algorithms like random forests performed better in analysis for large amount of data with complexity.

Reade, Singleton and Jewell (2020) in their study focussed on the initial pre-match coin toss and its outcome which influences decisions made by the playing teams. Similarly, Dawson et al. (2008) in their study looked into the potential advantage a team may have after winning the toss and their choice of batting order.

Cricket is being played for more than a century. There are plenty of data available relating to the game and that keeps on increasing with every match played. Consequently, there are numerous past studies and research work with diverse approaches and many more are being done regarding analysing cricket data and gathering insights.

## 5. Process Overview

The data analytical process in this project involved the following steps.

- **Data Collection:** First the data relating to the 2020 and 2021 seasons of the tournament was fetched mostly from cricsheet.org website and a few portions taken from Wikipedia. The data was in the form of comma-separated values(csv). Two sets of csv files were present. The first set contained a summary of each match with information on playing eleven, toss, winning team and winning margin. The second set contained the ball-by-ball data of every match that included the striker, non-striker, bowler, runs scored and wickets taken.
- **Data Cleaning and Preparation:** This phase was the most crucial and time consuming. The collected data in its raw form had to be cleaned and processed to be structured properly for analysis. This involved combining or segregating columns into usable data frame objects for analysing a particular field and was followed by removing duplicate rows and irrelevant columns and handling missing data. Sometimes new columns were also created based on existing ones.
- **Exploratory Data Analysis:** During this step multiple team-wise and player-wise parameters relating to batting and bowling in cricket were analysed. In this step observations were made that gave important insights like value range and trend or variation pattern of a parameter or interrelationship between certain parameters. The observation results were depicted in the form of colourful visualizations like pie charts, bar charts, scatter plots, histograms and network graphs. This stage helped in identifying overall team characteristics along with key players of every team.
- **Statistical Modelling using Machine Learning:** The goal of this stage was to attempt and find ways to predict the winning team of a match and the runs scored during the second innings of the teams. Different statistical approaches utilizing machine learning models were explored and the predicted results were compared with the actual ones to assess the performance of the models.

The analysis inclusive of visualization and modelling was primarily done using Python 3.8 programming language along with associated libraries like NumPy, Pandas, Matplotlib, Seaborn and Scikit-Learn. Microsoft Excel was used for some data investigation and preparation.

## 6. Exploratory Analysis Observations

### 6.1 Toss and Match Winners

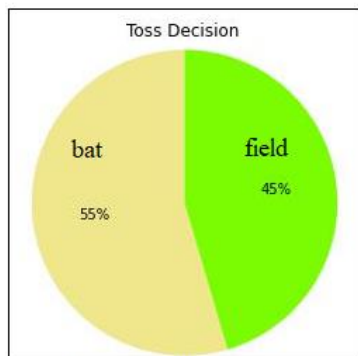


Figure 1a: Toss Decision

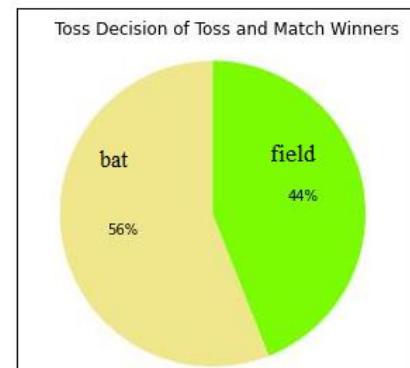


Figure 1c: Toss Decision of Toss and Match Winners

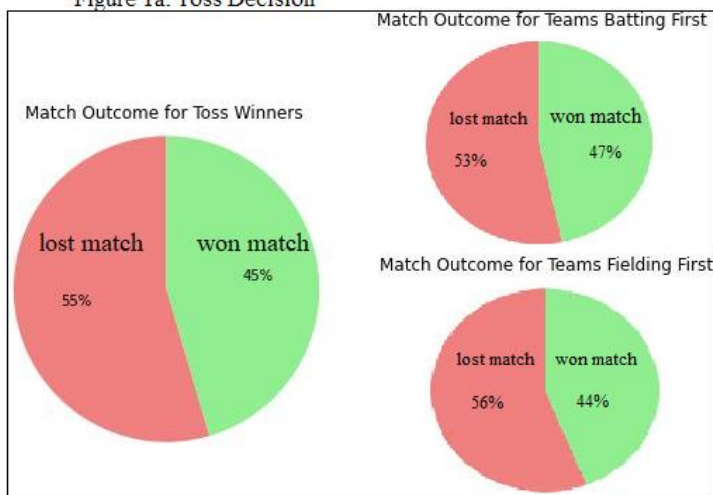


Figure 1b: Match Outcome of Toss Winners

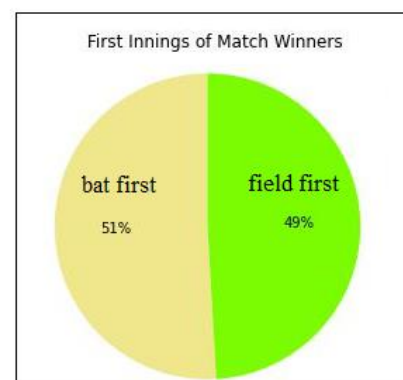


Figure 1d: First Innings state of Match Winners

The coin toss takes place before starting of every match between the captains of the two playing teams. The captain that wins the toss gets to decide whether to bat or field(bowl) first. This decision is often guided by factors such as the venue, the pitch and weather conditions and the areas of strength of the team. The captain winning the toss decides whatever would be advantageous for her team during the course of the match. For the matches played in the Rachael Heyhoe Flint Trophy, 55% of the time captains chose to bat first and the rest 45% chose to field first (*Figure 1a*). Teams that bat first, aim to score as many runs as possible in the first innings and the opponent is required to chase or exceed the run target of the first team in the second innings in order to win. Winning the toss however did not necessarily ensure that the team also won the match all the time. 45% teams won the toss as

well as the match while 55% toss winning teams lost the match. Teams choosing to bat first won the match 47% of the time and teams that chose to field(bowl) first won 44% of the time (*Figure 1b*). Of the teams that won both the toss and match majority (56%) chose to bat first and the remaining (44%) won the match after fielding in the first innings (*Figure 1c*). It has been observed that irrespective of winning or losing the toss, teams that eventually batted first appears to have a slight advantage over the teams that fielded first in winning the match. 51% of the matches were won by teams batting first and 49% were won by teams fielding first (*Figure 1d*).

## 6.2 Winning Teams and Winning Margins

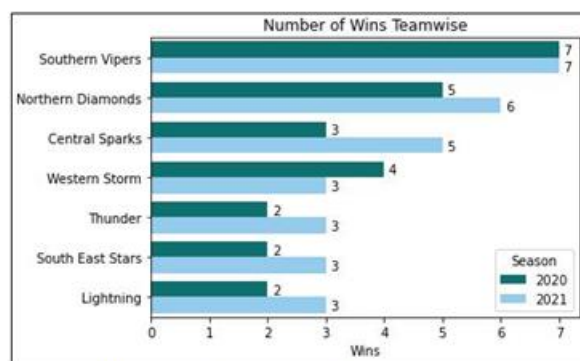


Figure 2: Number of wins teamwise

Among the eight teams, both times champions Southern Vipers have been the most successful team throughout the 2020 and 2021 seasons. They won all their matches in the 2020 season and won all except one match in 2021. The team has 14 wins out of 15 matches with 7 wins in each season. The other teams except Western Storm won more matches in 2021 than in 2020 season. Twice runners-up Northern Diamonds won 11 of 15 matches. Each of the teams North West Thunder, South East Stars and Lightning won 5 matches in total of which 2 were in 2020 and 3 were in 2021. The Sunrisers were the least successful team, having lost every match. (*Figure 2*)

When a team wins by defending their first innings run total, then the winning margin is represented by runs and when a team wins by successfully chasing the run target of the opponent then the winning margin is given in by wickets. Winning ‘by runs’ attributed to the bowling team, denotes the number of runs by which the team batting second falls short off reaching the required target. Winning ‘by wickets’ attributed to the second batting team, denotes the number of remaining wickets or batters left to bat for the team when they successfully score the required target runs. There have been victories of both great and narrow margins. Most matches have been won by 20 to 80 runs by the bowling team. There have been some close matches where the team managed to win by narrow run margins in the range of 1 to 10 runs. Some victories have also been achieved comfortably by high run margins of above 100 runs. For teams batting second, most wins came by 2 to 7 wickets. Seven matches have

been won by 6 wickets remaining. One match witnessed a close victory by 1 wicket while a few matches were comfortable victories for the chasing team with 8 and 9 wickets in hand. (Figure 3)

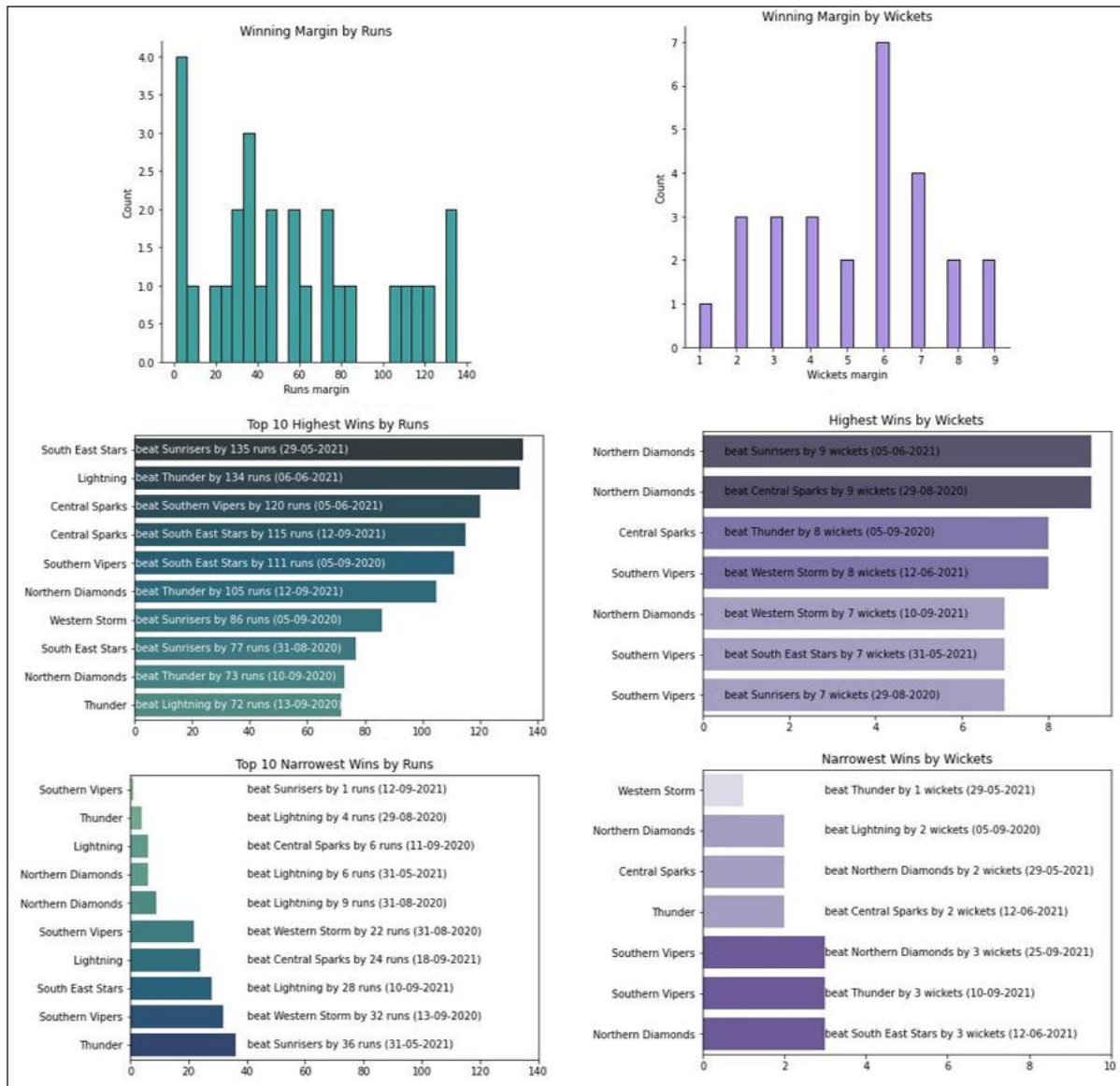


Figure 3: Winning margins by runs and wickets

### 6.3 First Innings Runs

It is always advantageous for the team batting first to put up as many runs as possible on the score board. This ensures a bigger target for the opponent to chase as well as gives more time and opportunities for the same team bowling second to take wickets and restrict the runs of the opponent. If the run target is higher, the team batting second will be under pressure to score enough runs quickly to keep up with the run rate. The players will have a greater tendency to attempt riskier batting shots to score boundaries. This in turn exposes the batters to more chances of getting out and once the wickets start to fall the pressure increases even more for the remaining batters. On the other hand, if



the target is reasonably low then the batters are more at ease and refrain from attempting reckless shots. The pressure in turn falls on the bowling team to take wickets quickly and limit as much runs as possible.

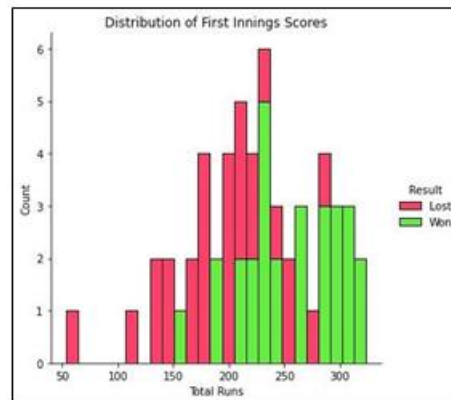


Figure 4: First Innings Runs Distribution

Most teams have scored within 180 to 300 runs in the first innings. Barring a few instances, teams having scored above 220 runs usually won the match. Similarly, except for a few occasions, teams scoring less than 200 runs usually lost. (Figure 4) In the 2021, there were four matches where runs above 300 were scored in the first innings while in the previous year this happened only once. The highest runs scored was 324 by South East Stars against Sunrisers on 29th May 2021. The highest first innings runs in 2020 season was 303 by Lightning against Central Sparks on 19th September 2020. The highest runs total successfully chased, was 291 by Western Storm who scored 295 losing 9 wickets and thus winning against North West Thunder by 1 wicket on 29th May 2021. Sunrisers were bowled out for only 53, the lowest ever runs in the first innings, against Northern Diamonds on 5th June 2021. The lowest runs total of 151 was successfully defended by Northern Diamonds versus Lightning who were bowled out for 145 resulting in Northern Diamonds winning by 6 runs on 31st May 2021. (Figure 5)

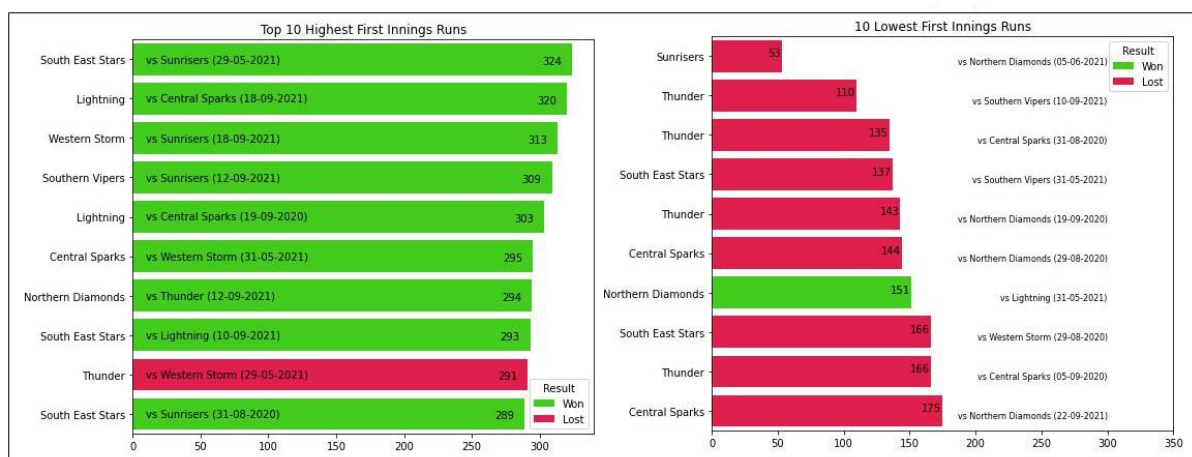


Figure 5: Highest and Lowest First Innings Runs

## 6.4 Individual Batter Runs

Every team's playing eleven consists of designated batters and allrounders who are given priority in the batting line up of each team. They are expected to score most of the runs for the team. When the batters and allrounders get out then the bowlers have to go out to bat. Usually, bowlers are not expected to score as many runs as the batters or allrounders. When individual batters and allrounders score more runs not only does their individual figures improve, but this also puts their team in a more strategic position in the match.

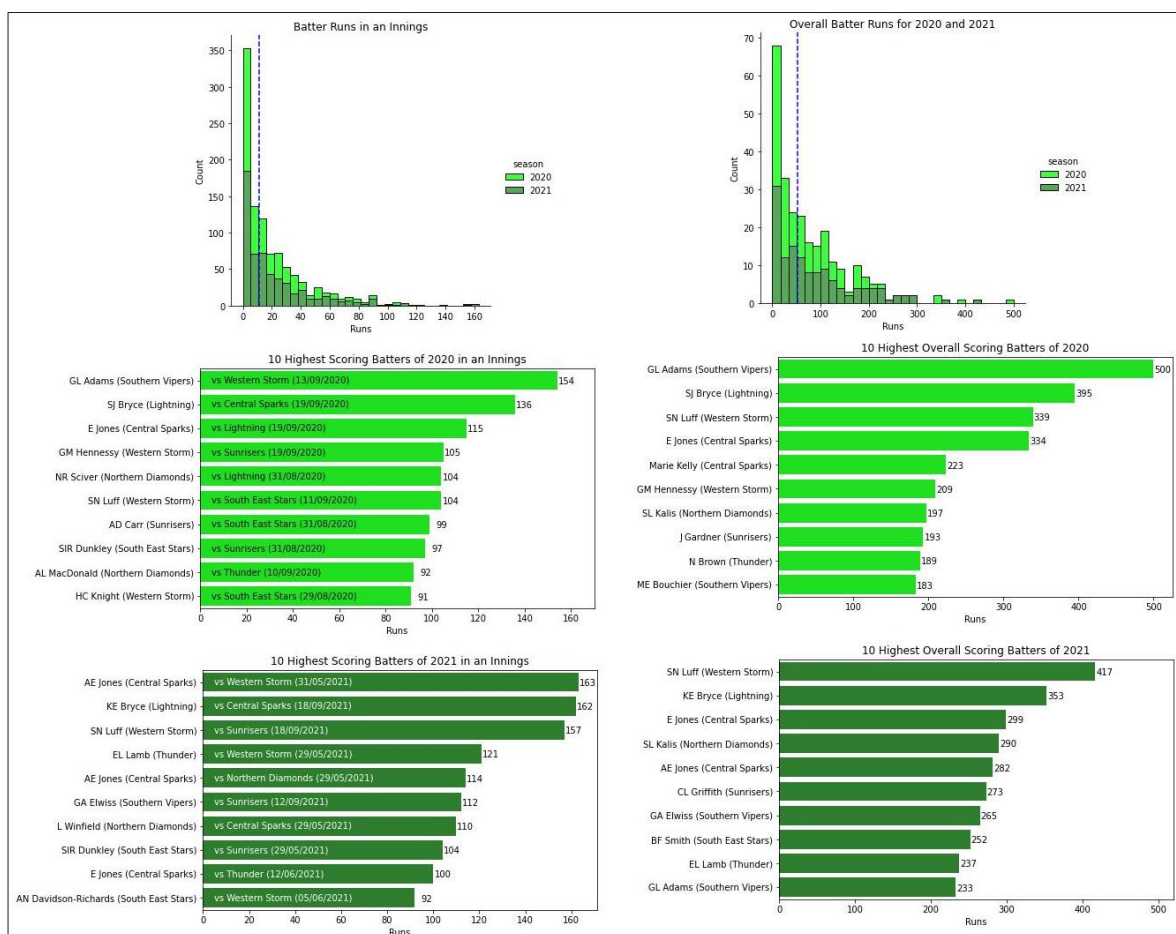


Figure 6: Individual Batter Runs in an Innings and Overall

Over the two seasons of the Rachael Heyhoe Flint Trophy, the median runs scored in a single match by an individual player inclusive of being a batter, allrounder or bowler, was around 10. Most players scored between 0 to 40 runs in a match. Comparatively a smaller number of players scored 50 runs or more with only few reaching the coveted milestone of scoring a century or more runs. The highest individual runs scored in 2020 was 154 by GL Adams from Southern Vipers against Western Storm on 13<sup>th</sup> September. The second highest runs scored was 136 by SJ Bryce from Lightning against Central Sparks on 19<sup>th</sup> September. Four other batters scored centuries in that year. Two players from Western Storm SN Luff and GM Henessey scored 104 and 105 respectively. E Jones from Central Sparks scored 115 and NR Sciver from Northern Diamond scored 104. More centuries were scored in

the year 2021. Every team except Sunrisers had at least one player scoring century. AE Jones from Central Sparks scored centuries twice, one being 114 against Northern Diamonds on 29<sup>th</sup> May and other was 163 against versus Western Storm and on 31<sup>st</sup> May. The latter was also the highest individual runs of 2021. Central Sparks also had another century scorer E Jones with 100 runs. KE Bryce from Lightning scored 162 followed by SN Luff from Western Storm 157 and EL Lamb from Thunder with 121. For Southern Vipers, GA Elwiss was the highest individual run scorer with 112. For Northern Diamonds it was L Winfield with 110 runs and similarly for South East Stars it was SIR Dunkley who scored 104 runs. In case of cumulative runs scored in all the matches in each year by players (inclusive of batters, allrounders or bowlers), the median was around 50. Most had a net run tally ranging from 0 to 100 while some scored above 200 and few even scored more than 300 runs. GL Adams from Southern vipers was the highest net run scorer of 2020 season with 500 runs from all the matches. SJ Bryce from Lightning was second highest with 395 runs followed by SN Luff from Western Storm with 339 and E Jones from Central Sparks with 334. In 2021, SN Luff had the highest net runs of 417. Only another batter KE Bryce from Lightning had above 300 runs with 353.

(Figure 6)

## 6.5 Batting Strike Rate

Batting Strike Rate is a measure of the average number of runs scored by a batter per 100 balls faced. (Wikipedia., 2022b)

$$\text{Batting Strike Rate} = \frac{\text{Runs}}{\text{No. of balls faced}} \times 100$$

Batting Strike Rate helps to understand how quickly a batter is scoring runs and higher value is desirable especially in limit overs cricket matches. A strike rate above 100 implies that runs scored by the batter is greater than the number of balls taken to score. However, strike rate solely does not indicate the batting skills of a player. A player may score low runs and yet have a high strike rate if she takes lesser number of balls to score.

Over the two years of the tournament, most batting strike rates in an innings varied between 0 and 100 with the median being around 50. In general, most low batting strike rates were observed for those who scored less than 20 runs. The strike rate rarely exceeded 150 for some batters scoring runs in the range of 0 to 49. Batters with above 50 runs had strike rates varying between around 60 and around 150. The range was closer to 100 for batters scoring centuries. (Figure 7a)

For overall comparison throughout all the matches in both seasons, the mean of the strike rates of each player was taken and their distribution observed. The median value was around 55. For batters with match experience from 1 to 5, their mean batting strike rates ranged from 0 to 100 with few exceeding 120. The range becomes narrower from around 30 to 100 for batters who have played between 6 to 10 matches. The most experienced batters having played over 11 matches have mean

batting strike rates ranging from around 40 to 80. (Figure 7b)

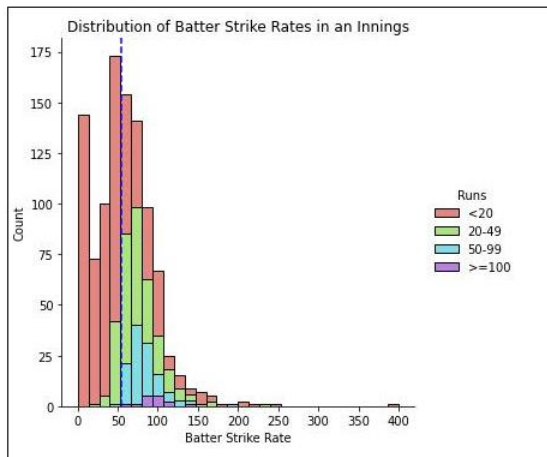


Figure 7a: Distribution of Batting Strike Rates in an Innings

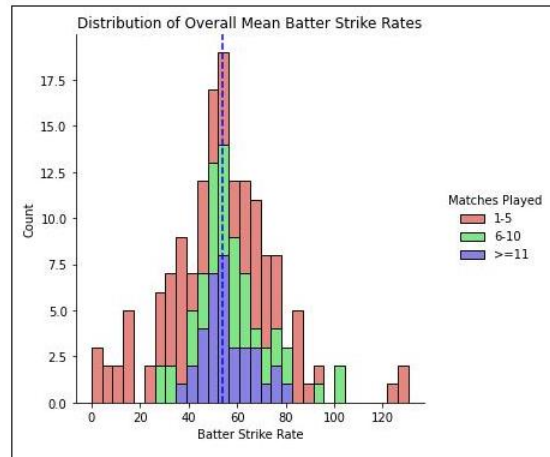


Figure 7b: Distribution of Mean Batting Strike Rates

## 6.6 Batting Average

Batting Average is a performance parameter of batters given by the ratio of the total runs they have scored to the number of times they have been out. (Wikipedia., 2022c)

$$\text{Batting Average} = \frac{\text{Total Runs}}{\text{No. of dismissals or outs}}$$

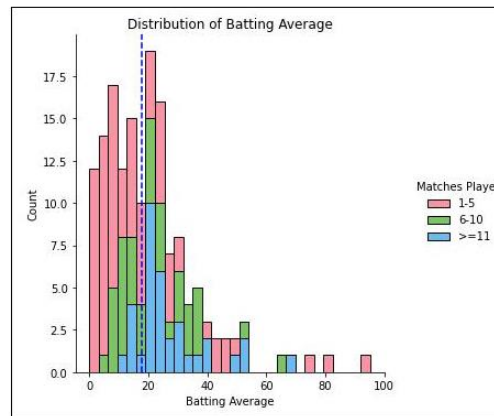


Figure 8: Distribution of Batting Average

Higher batting average indicates better consistent batting performance. The median batting average considering matches of the two seasons was close to 20. Most batters had batting averages within the range from 0 to 40 while only some had above 40 along with few above 60. For players with higher experience of more than 11 matches, their batting average ranged from around 10 to 40 with only few beyond that. Most lower batting averages were of players with match experience of 1 to 5 barring a few exceptions whose were above 40. (Figure 8)

## 6.7 Batting Partnerships

During batting, two players- the striker and non-striker from the team are present on the pitch and form a batting partnership and score runs in tandem. The partnership continues until one of them gets out. When a new player comes to bat, a new partnership is created. One player can be part of multiple partnerships which usually happens if the player manages to bat for a long time while her partners get out. Partnership runs are calculated by considering the runs scored for the team when those two batters were on the ground together. It is always desirable for a batting team to have stable partnerships where both batters are scoring decent runs without getting out.

In 2020, there were 18 times when batting partners scored more 100 or more runs. GM Hennessy and SN Luff from Western Storm together scored 162, the highest ever partnership runs of the season against South East Stars on 11th September. In comparison the 2021 season had 13 occasions where 100 or more runs were scored in partnerships. However, the highest partnership runs in that season exceeded 200. KE Bryce and SJ Bryce from Lightning scored 207 together versus Central Sparks on 18th September. When considering the overall net runs accumulated in partnerships, it was found that GL Adams and EM McCaughan from Southern Vipers scored 407 runs between themselves in all the matches of 2020. Five other partner pairs scored above 200 runs of which two were from Southern Vipers, two from Central Sparks and one from Lightning. In the following year, eight partner pairs scored above net runs together above 200. Two were from Sunrisers and there was one each from Central Sparks, Lightning, Western Storm, South East Stars and Southern Vipers. The highest was Marie Kelly and E Jones from Central Sparks with 281 runs. (*Figure 9*)

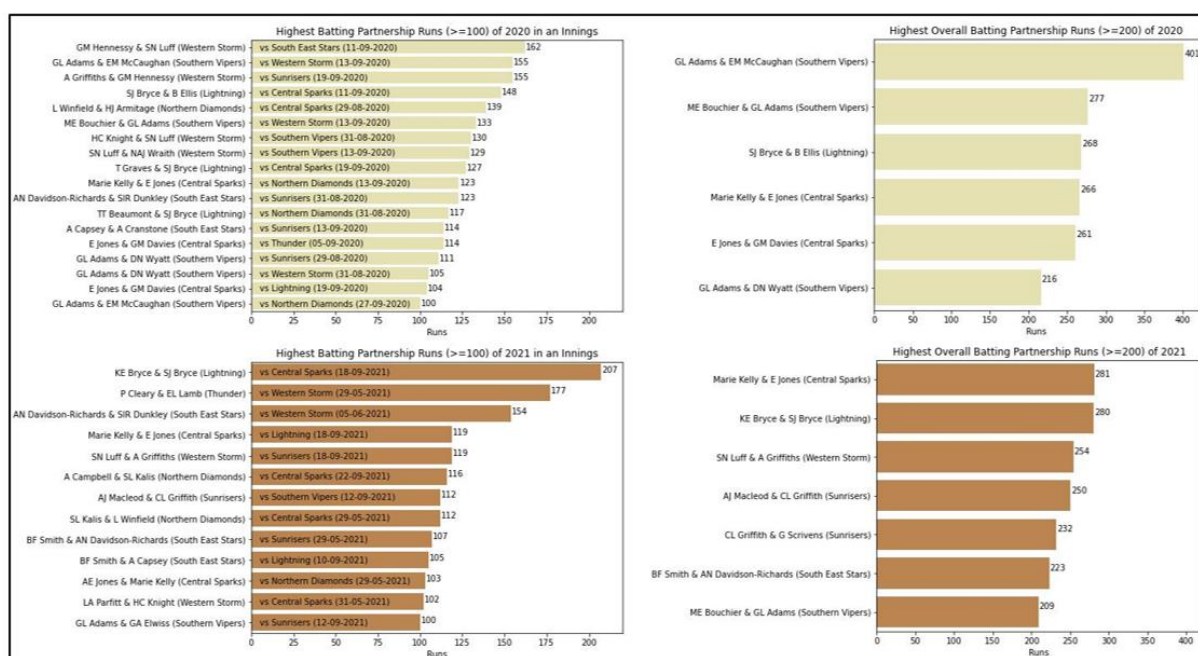


Figure 9: Partnership Runs in an innings and overall

### 6.7.1 Partnership Networks of different teams and runs scored

Taking reference from the work of Mukherjee (2013), the batting partnerships of the different teams have been studied using a network graph representation. Each node in the network represents each batter while an edge or a connection between them denotes a partnership. The direction of the arrows goes towards the dominating partner or the batter who contributed more runs in that partnership. The distribution of net runs scored in the two years by each pair of batting partners of each team have also been studied.

**Southern Vipers** - The Southern Vipers had 16 different players batting for them in the two years of the tournament and there were 48 unique partnerships among those players. GA Elwiss and E Windsor participated in partnerships with 11 other players followed by CE Dean with 9 partnerships and GL Adams with 8. In terms of scoring the higher proportion of runs in the partnerships, CE Dean scored more runs in 7 of her 9 partnerships. E Windsor was the dominant run scorer in 7 out of 11 partnerships. GA Elwiss scored more in only 4 out of 11 partnerships and for GL Adams it was 4 out of 8. The overall runs scored by the batting partners for Southern Vipers ranged from 0 to around 120 with the median being around 25 to 30 runs. High scoring partnerships were generally those playing more matches with few having above 150 and some even close to 400 and 500 runs. (*Figure 10*)

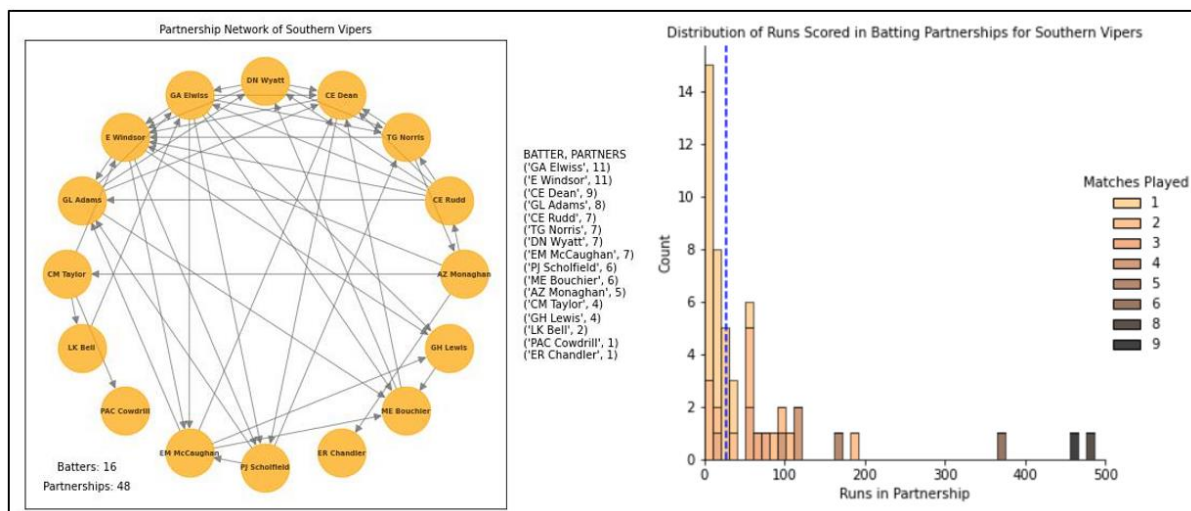


Figure 10: Partnership Overall Network and Net Runs of Southern Vipers

**Western Storm** - For Western Storm there were 21 different batters participating in 50 different partnerships. SN Luff partnered with 13 other players and was the higher run scorer in 7 of them. NAJ Wraith scored higher in 5 out of her 9 partnerships and DR Gibson in 5 out of her 8. Most partnerships played only one, two or three matches and the runs scored by them ranged from 0 to 60. The median was close to 20. Some partnerships with more match experience scored above 100 net runs with few being close to 200 and 300 and the highest was around 350. (*Figure 11*)



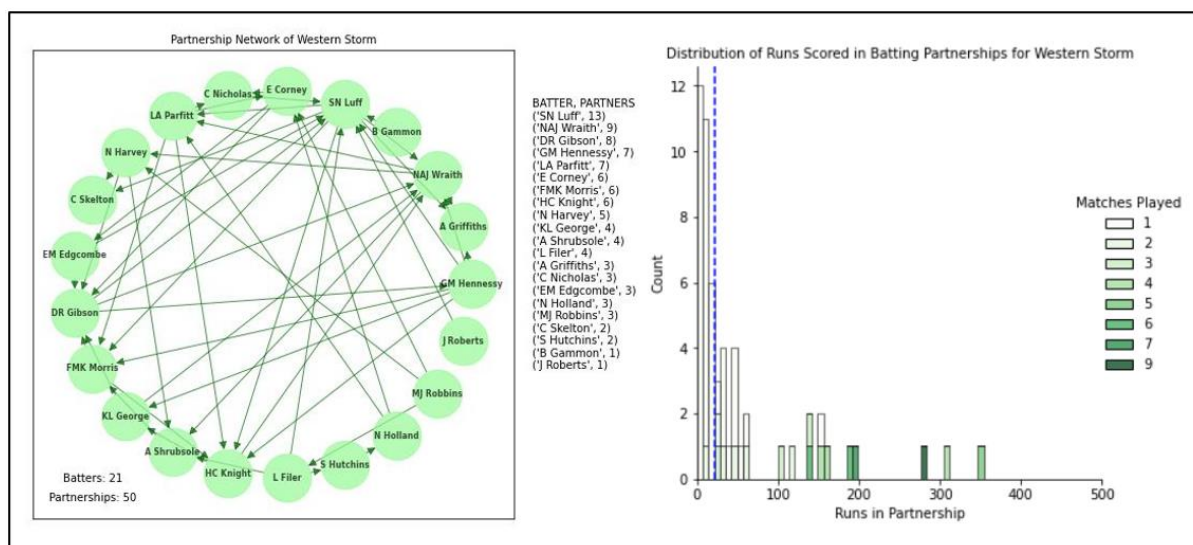


Figure 11: Partnership Overall Network and Net Runs of Western Storm

**Lightning** - The team had 53 different partnerships among 22 players. AJ Freeborn was in 12 partnerships followed by KE Bryce in 11. SJ Bryce and T Graves both were part of 9 partnerships. Out of 12, AJ Freeborn was the dominant run scorer in 8. KE Bryce also scored more runs in 8 of her 11 partnerships. Both SJ Bryce and T Graves scored higher in 5 out of 9 partnerships. Most partnerships with less match experience scored runs in the range of 0 to 80 with the median being close to 20. Some batting partners having played more than four matches scored more than 100 runs with a few having close to 200 and 300. The highest net runs scored by a partnership with match experience of eight was close to 430. (Figure 12)

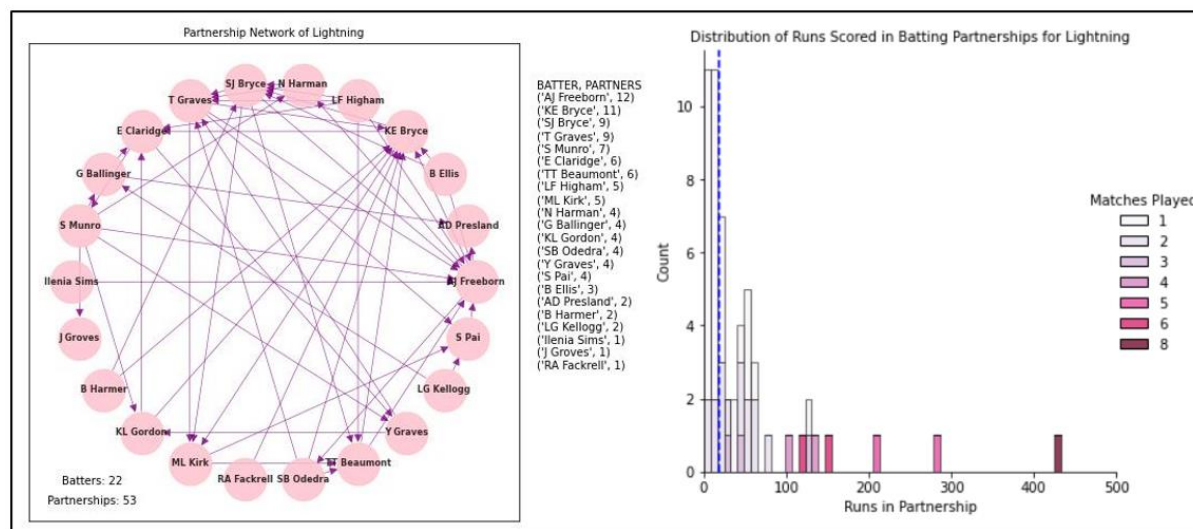


Figure 12: Partnership Overall Network and Net Runs of Lightning

**Northern Diamonds** - The number of unique batting partnerships for Northern Diamonds was 58 and 20 different batters were involved. SL Kalis was part of 13 partnerships and her run contribution was higher in 7 of those. A Campbell was in 11 partnerships, and she scored more runs in 8 of those. BAM Heath scored more runs in 4 out of her 9 partnerships. The net runs scored by the different

partnerships ranged from 0 to 90 with the median being around 30 runs. The higher run scoring partnerships in the range of 150 to 280 runs were those having played five or more matches.

(Figure 13)

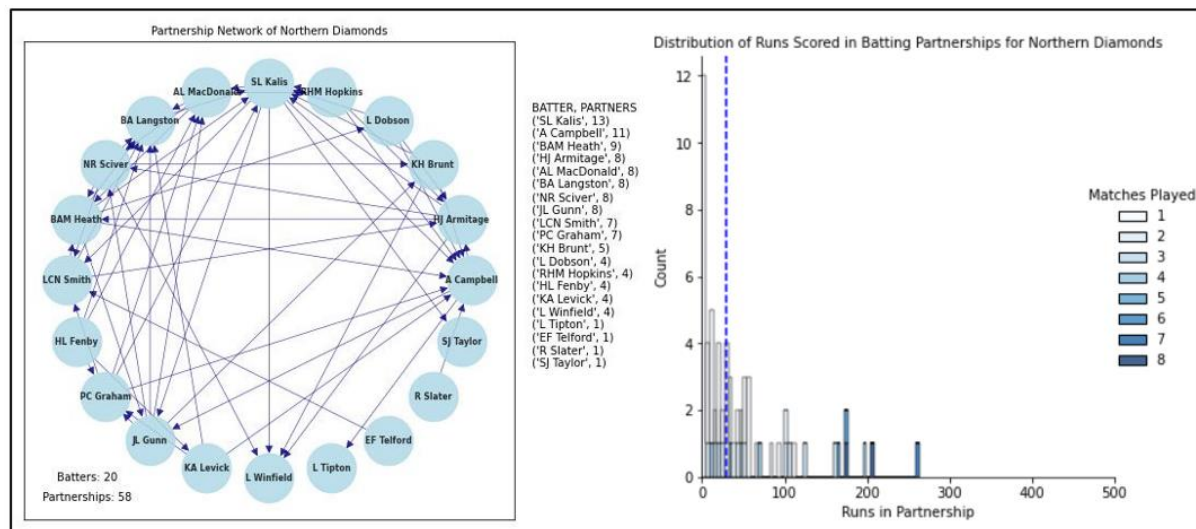


Figure 13: Partnership Overall Network and Net Runs of Northern Diamonds

**Sunrisers** - The Sunrisers had 18 players and 60 different partnerships. KS Castle and J Gardner each were partnered with 12 other players with the former contributing more runs in 4 of them while the latter scored more runs in 5 of them. AD Carr, G Scrivens and KL Midwood each were part of 10 partnerships. AD Carr scored more runs in 7 out of 10, G Scrivens in 5 and KL Midwood in only 3 out of 10. The net runs scored for Sunrisers by all the partnerships ranged from 0 to 90 with the median being around 15 runs. Few partnerships accumulated more runs within 150 to 200 while there were some who had more than 200 runs with the highest runs scored being even above 300.

(Figure 14)

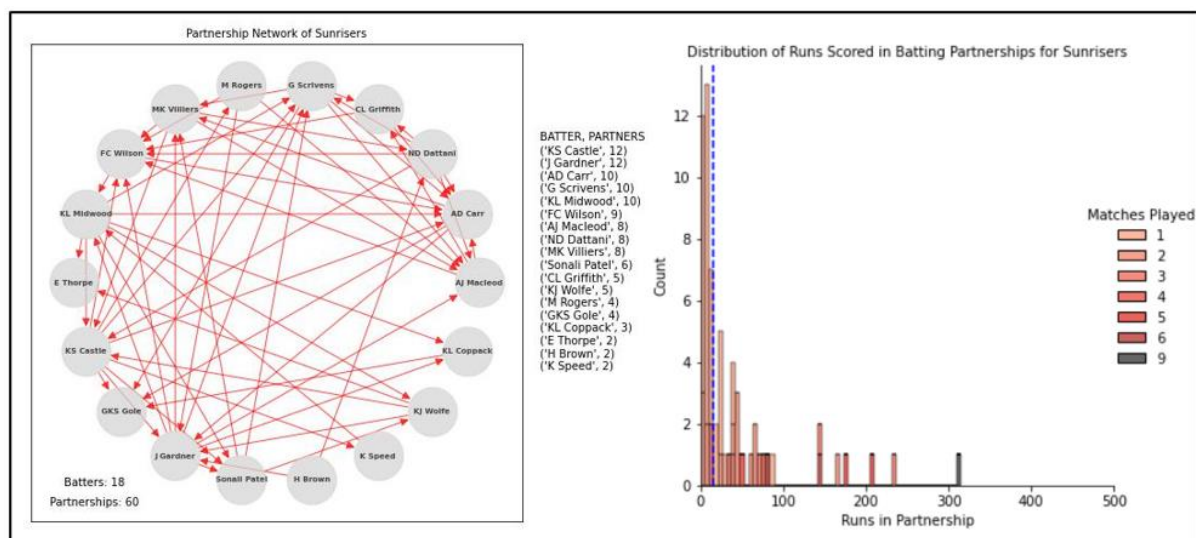


Figure 14: Partnership Overall Network and Net Runs of Sunrisers



**Central Sparks** - There were 62 unique partnerships for Central Sparks with 19 players involved. E Jones was part of 15 partnerships and her presence in different partnerships was not only the highest in her team but also among other teams. She contributed more runs in 9 of them. CAE Hill was in 10 partnerships, and she scored more runs in 8. GM Davies, Marie Kelly and IECM Wong all were in 9 partnerships. GM Davies scored higher in 7, Marie Kelly in 6 and IECM Wong in only 1. The net runs in partnerships ranged from 0 to 100 with 25 being the median and the most partnerships played only up to three matches. The exception was only one partnership that played nine matches and scored close to 380 runs. (*Figure 15*)

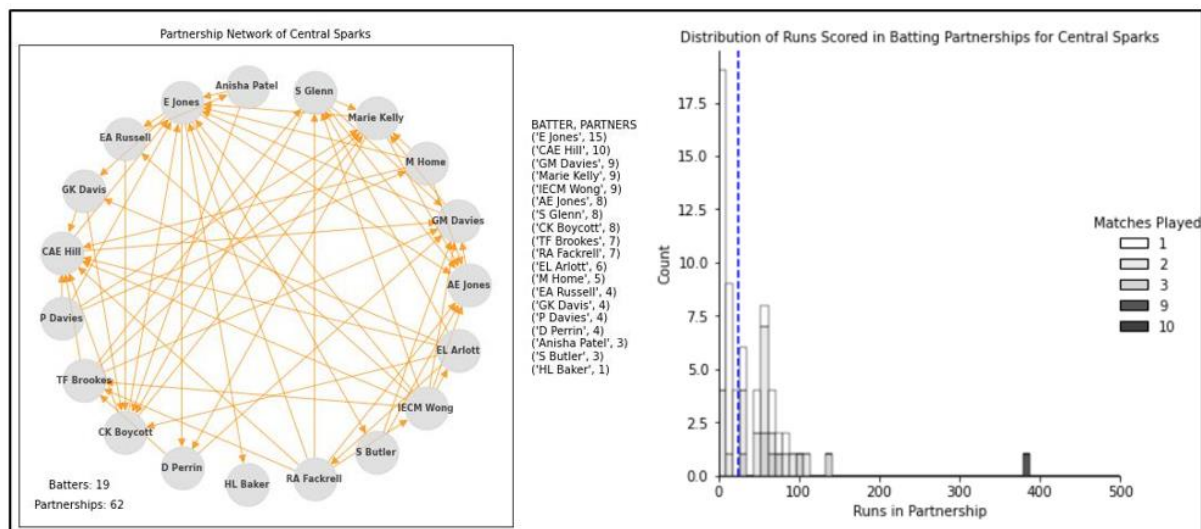


Figure 15: Partnership Overall Network and Net Runs of Central Sparks

**North West Thunder** - North West Thunder had 63 partnerships. Of the 20 different players who batted, E Threlkeld was in 13 partnerships, contributing more runs in 7 of them. KL Cross played in 12 partnerships, scoring more runs in 9 of them. GEB Boyce scored more runs in 6 of her 11 partnerships and N Brown in 7 of her 10. Most batting partnerships scored net runs ranging from 0 to 90 with median between 20 to 25 in the matches played in the two seasons. Few partnerships had net runs in the range from 100 to 200. (*Figure 16*)

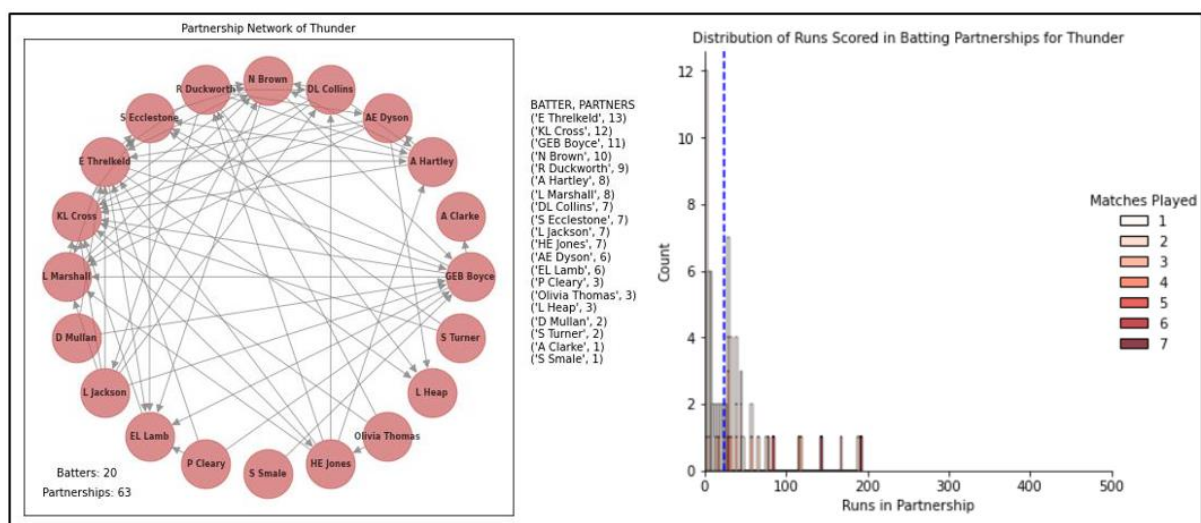


Figure 16: Partnership Overall Network and Net Runs of Thunder

**South East Stars** - South East Stars had 70, the highest number of unique batting partnerships as well as 23, the greatest number of different players batting for the team. The players with most batting partners were A Cranstone and A Capsey each of them being part of 10 partnerships. A Cranstone scored the higher proportion of runs in 3 partnerships while A Capsey scored in 5 of them. The net run range was between 0 to 80 and the median was between 15 to 20. There were few experienced partnerships that went on to score above 100 and also 200 runs. The highest was close to 300 runs. (Figure 17)

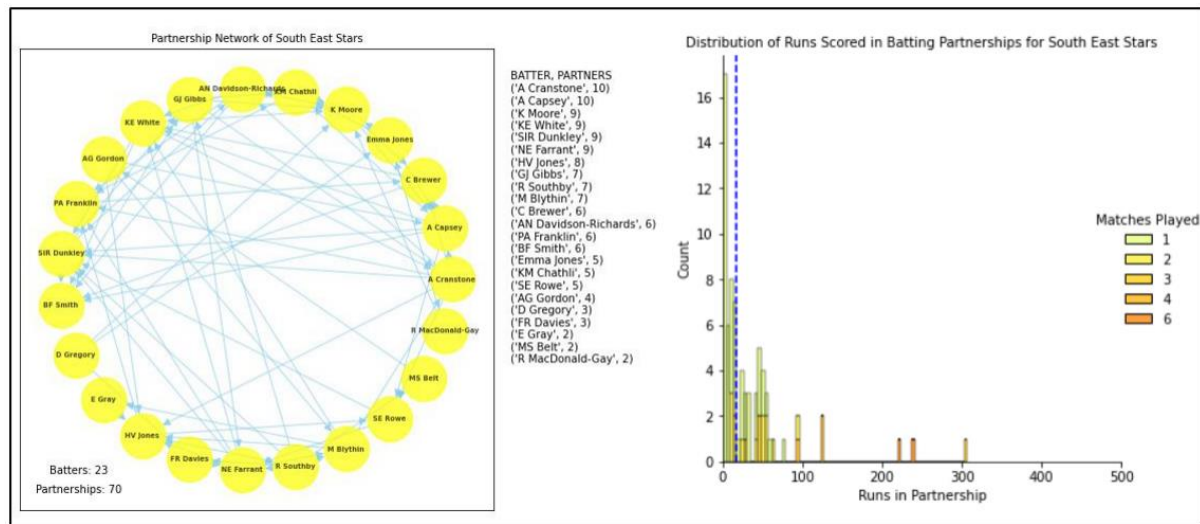


Figure 17: Partnership Overall Network and Net Runs of South East Stars

It was noticed that the distributions of net runs by the partnerships of all the teams are synonymous with each other where majority of the runs are scored in the range of 0 to 100 and the median lies from close to 15 to within 30. It was only those partnerships scoring 100 and more runs throughout both seasons of the tournament, which have made significant impact in the batting performances of the teams.

## 6.7.2 Partnership Networks and Runs of Southern Vipers and Sunrisers per match

In order to understand the impact of batting partnerships per match, the partnership networks of the two teams on the extreme positions of success - the Southern Vipers and the Sunrisers have been observed.

For most of the matches, it was not required for all the eleven players of Southern Vipers to come to bat and the team could finish the innings with wickets in hand. Lesser number of partnerships indicated a stable batting performance of the team because such happens when most part of the run scoring is handled by the batters and allrounders of the team with less dependence on the bowlers. As shown in the below *Figure 18*, the Southern Vipers won eight of their fourteen matches batting second and in three of those they could manage to chase the target with 6 partnerships. In two matches 4 partnerships were sufficient to reach the target and for another one it was just 3. One match required 7 and another one required 8 partnerships. Of the six matches won batting first, two of them had 8 and

another two had 9 partnerships. One had only 2 partnerships while only one had 10 partnerships requiring all the eleven players to bat. They only lost one match (shown in red background) where they had 5 partnerships with 6 batters.

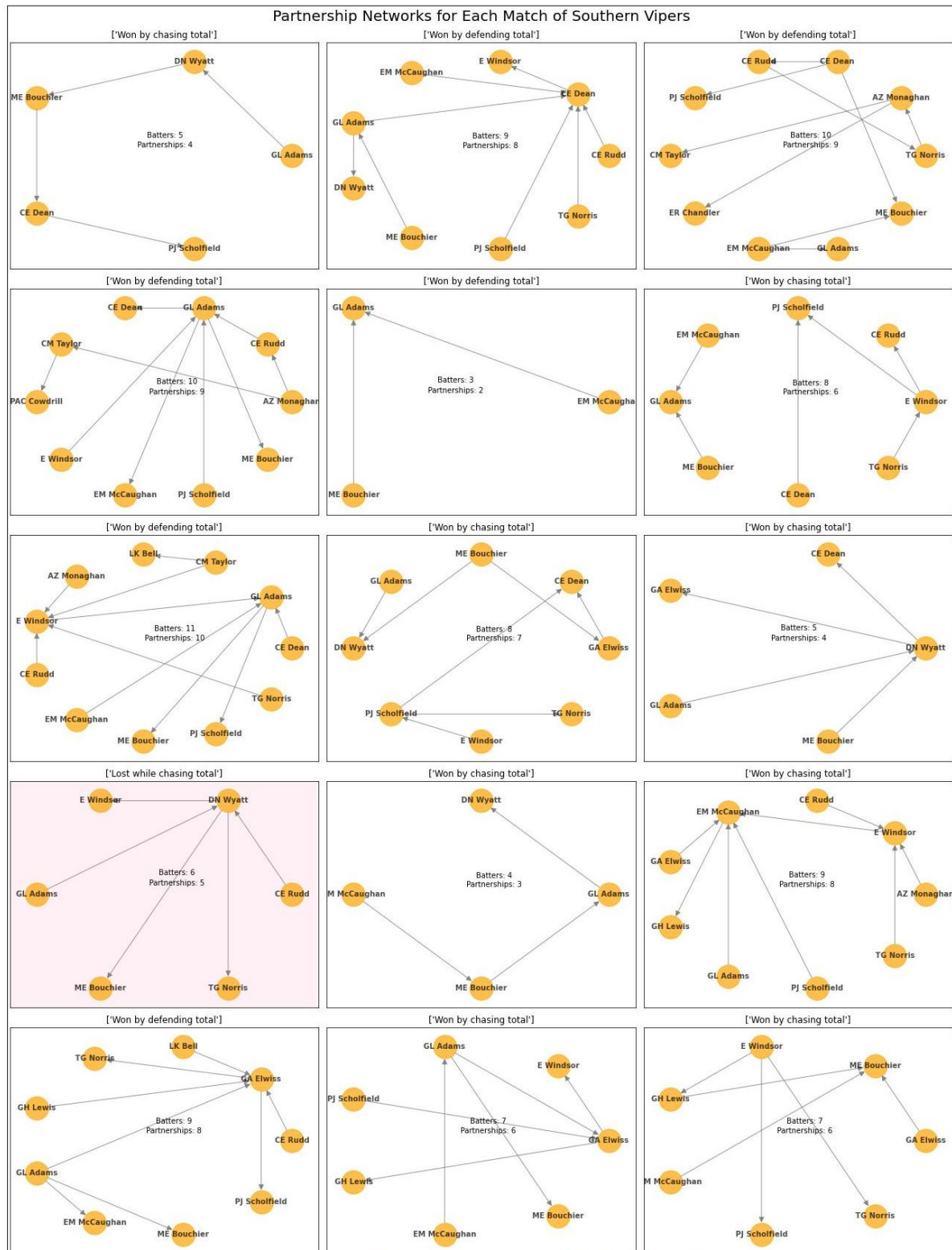


Figure 18: Partnership Network of each match of Southern Vipers

The Sunrisers lost all their matches and in six of those matches, they needed 9 partnerships while batting and in two of those, all eleven players had to come out to bat. For two matches, they needed 10 partnerships and then also all the eleven players had to bat. Three matches had 8 while two had 7 partnerships. (Figure 19)

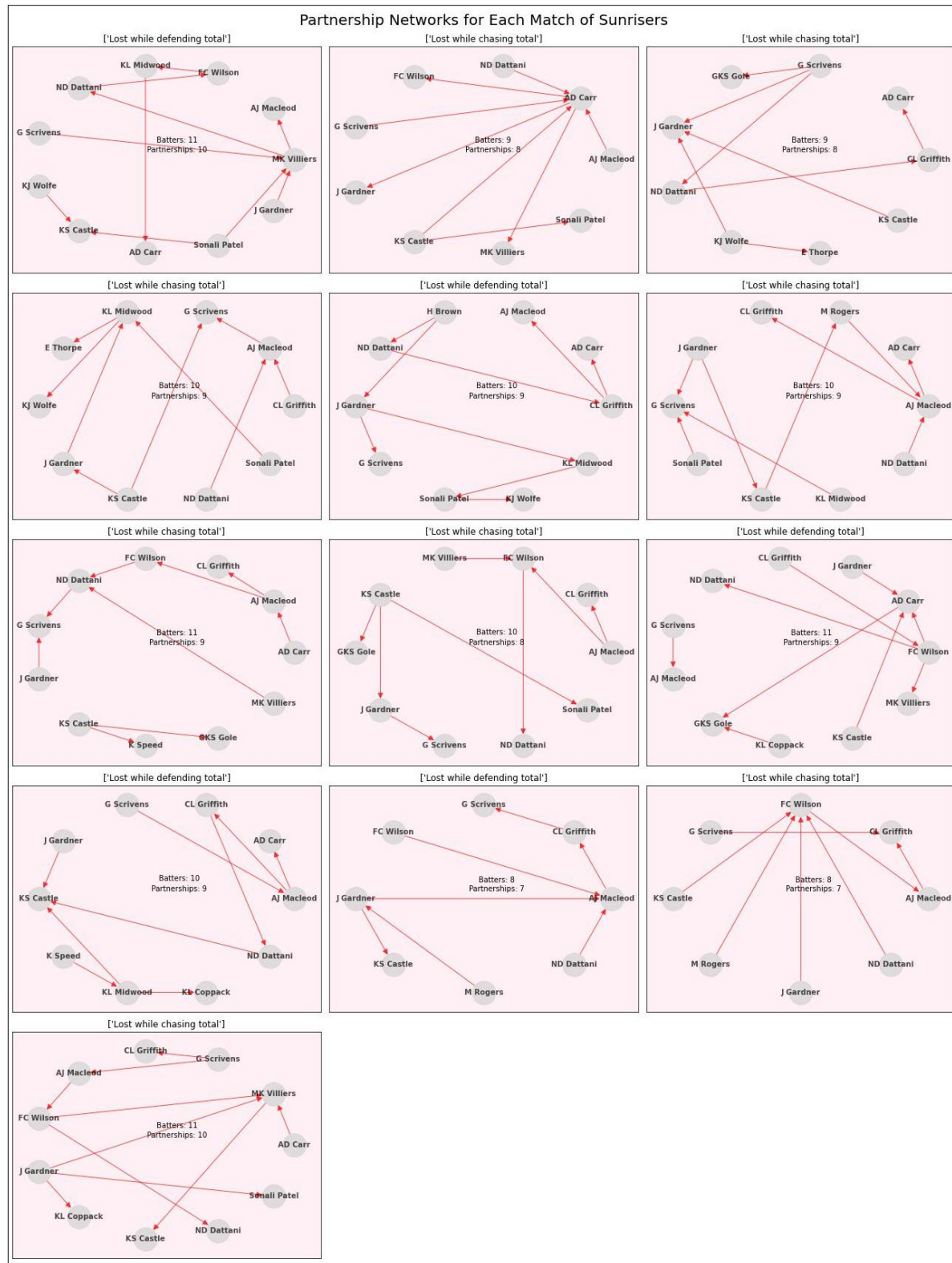


Figure 19: Partnership Network of each match of Sunrisers



The median runs scored by batting partnerships in an innings of both Southern Vipers and Sunrisers lie within 15 to 20. However, in the upper half region beyond the median, the batters of Southern Vipers scored more runs than the batters of Sunrisers particularly in the 20 to 60 range. Also, there are a greater number of partnerships for Southern Vipers who have scored 100 or more runs compared to Sunrisers. (Figure 20)

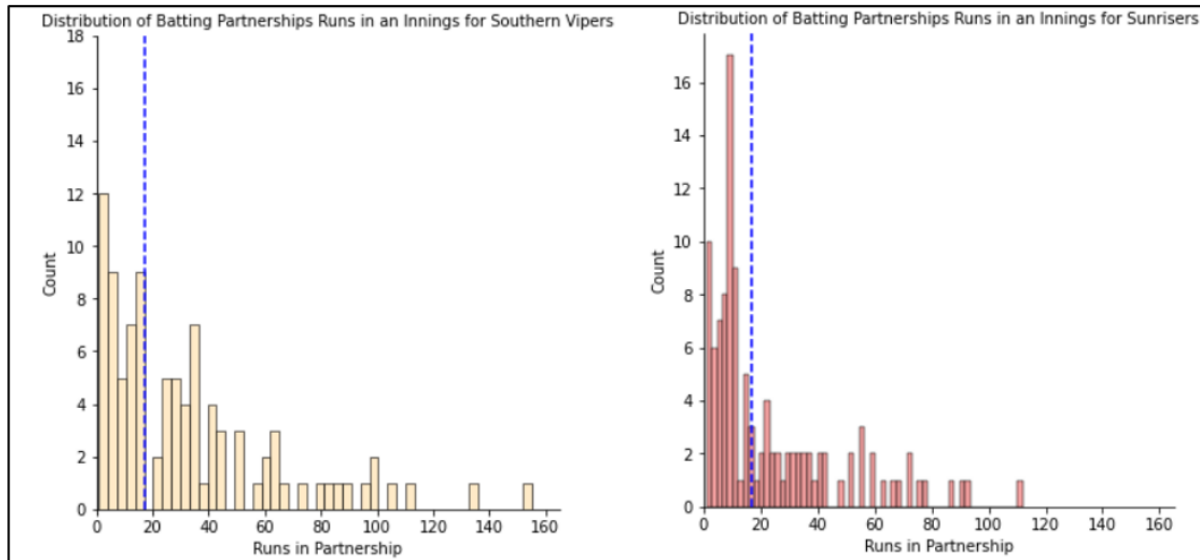


Figure 20: Distribution of Partnership Runs in an Innings of Southern Vipers and Sunrisers

## 6.8 Wickets and Wicket-takers

Bowlers aim to take wickets to impede the run scoring of the opponent. The batting innings of the opponent gets concluded prematurely once the bowlers manage to take 10 wickets implying that all opponent batters have been out. The batters can get out when the ball manages to hit the wicket sticks behind the batter. Other than that, there are other different ways of getting out and these are also denoted by the term 'wickets'. Most type of wickets except for 'run out' is credited to the current bowler when the wickets are taken. (Marylebone Cricket Club, 2017)

Throughout the two seasons of Rachael Heyhoe Flint Trophy, 858 wickets have been taken. The most common way of batters getting out was by the ball being caught by a fielder before reaching the ground. The next common way of getting wickets was to bowl out the batter by bowling straight to the wicket sticks behind the batter. LBW (Leg Before Wicket) accounted for around 17% of the wickets. This happens when the batters leg obstructs the flow of the ball which would have hit the wickets on the pitch. 8% of the time the batters were 'run out'. Stumpings, getting 'caught and bowled' and 'hit wicket' happened less frequently. (Figure 21) 'Caught and bowled' refers to the situation when the ball is caught by the bowler herself before reaching the ground. A batter is 'stumped' out when the opponent wicket keeping fielder displaces the wicket sticks with the ball in hand while the batter has moved out from her position on the ground. 'Hit wicket' occurs when a

batter mistakenly displaces the wicket sticks behind her by herself or with the bat while facing the ball from the bowler.

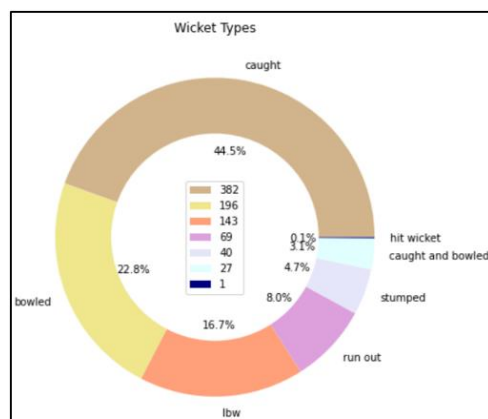


Figure 21: Wicket Types

Like runs for a batter, the number of wickets is an important performance parameter for a bowler. The difference however is that unlike runs which can increase to any extent, in one match the bowlers can at most take 10 wickets and one bowler is only allowed to bowl a maximum of 10 overs or 60 balls in a 50 over game. So, it is generally challenging for one bowler to take more than 1 wicket in a single match. In the tournament during a match 221 bowlers have taken 1 wicket while 133 have managed to take 2. Only 56 could take 3. Just 22 bowlers have taken 4 wickets in a single match. A 5-wicket haul was achieved by 8 bowlers; 3 of them in 2020: KE Bryce from Lightning, FMK Morris from Western Storm and KH Brunt from Northern Diamonds. The other 5 achieved this in 2021: HE Jones from Thunder, IECM Wong, EL Arlott from Central Sparks, LCN Smith from Northern Diamonds and NE Farrant from South East Stars. The incredible feat of taking 6 wickets single-handedly was done in the final match of 2020 season by CM Taylor from Southern Vipers against Northern Diamonds on 27<sup>th</sup> September. (Figure 22)

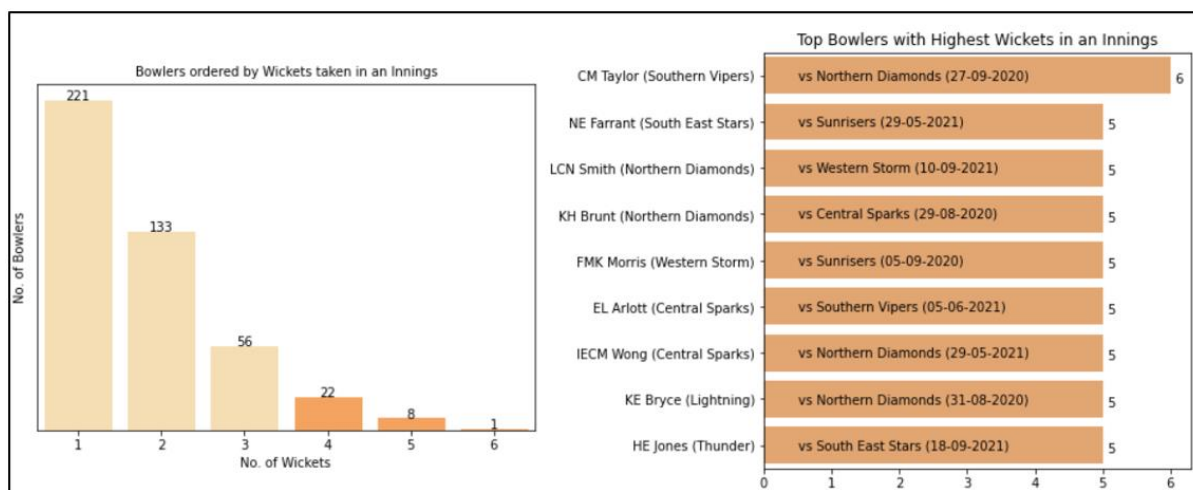


Figure 22: Wickets Taken by Bowlers in an Innings

CM Taylor was also the highest wicket-taker with 15 wickets in 2020. Only seven other bowlers took 10 or more wickets in that year. In the following year, the number of bowlers taking 10 or more

wickets increased to sixteen. KL Gordon from Lightning was the highest wicket-taker in 2021 with 16 wickets. (Figure 23)

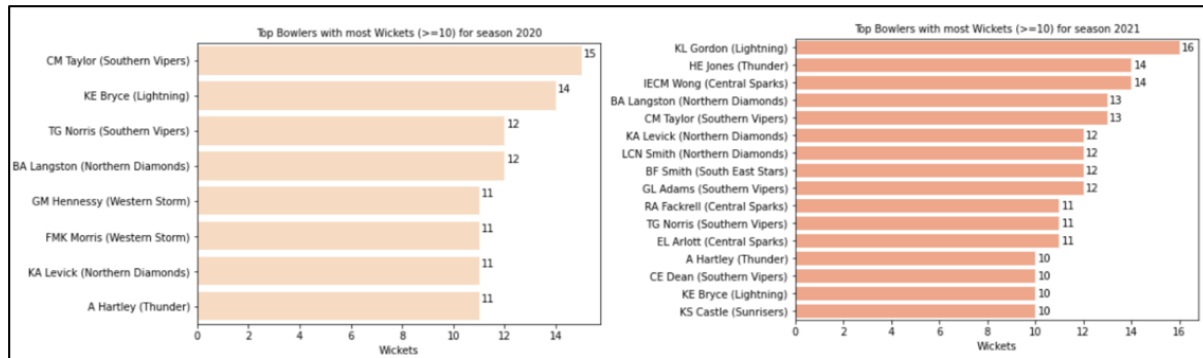


Figure 23: Highest Wicket taking Bowlers

## 6.9 Bowling Economy Rate

The economy rate of a bowler is the measure of the number of runs, a bowler concedes in an over. (Wikipedia., 2022d)

$$\text{Economy Rate} = \frac{\text{Runs Conceded}}{\text{Overs Bowled}}$$

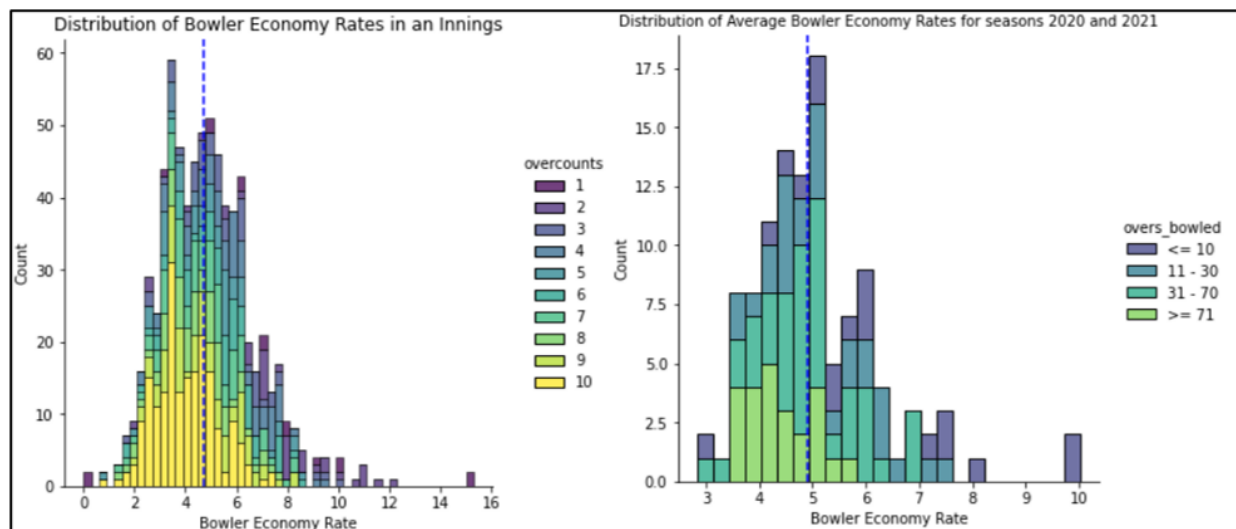


Figure 24: Economy Rate in an Innings and Overall

Bowlers prefer to avoid being 'expensive' or to have a low bowling economy rate as it implies that opponent batters have scored less runs in those overs. The economy rate is increased when a batter scores more runs in an over and also when a bowler bowls 'extras' like 'wide' and 'no-ball' which awards runs to the opponent without that ball taken into consideration. Most bowlers in one match had economy rates in the range from 2 to 8 and the median was near 4.5. For some bowlers who have bowled up to three overs, the economy rate was on the higher side, above 8. For the majority of bowlers who have bowled seven or more overs, the economy rate was on the lower side of the median. When the economy rate was studied throughout all the matches played in both seasons, it was found that bowlers with greater experience of bowling more than seventy-one overs had rates in the range from around 3.5 to 5.8. The median overall economy rate was close to 5. For few bowlers having bowled thirty overs or less, the overall economy rate exceeded 7. (Figure 24)

## 6.10 Bowling Strike Rate and Bowling Average

For wicket taking bowlers, bowling strike rate and bowling average are two other performance parameters. Bowling strike rate gives an indication of how quickly a bowler has taken wickets. (Wikipedia., 2022e) Bowling average helps to understand how expensive the bowler was while taking wickets. (Wikipedia., 2022f)

Like economy rate, lesser values of these two parameters are desirable. Bowlers taking more than one wicket usually have lower values of bowling strike rate and bowling average.

$$\text{Bowling Strike Rate} = \frac{\text{Balls Bowled}}{\text{Wickets}}$$

$$\text{Bowling Average} = \frac{\text{Runs Conceded}}{\text{Wickets}}$$

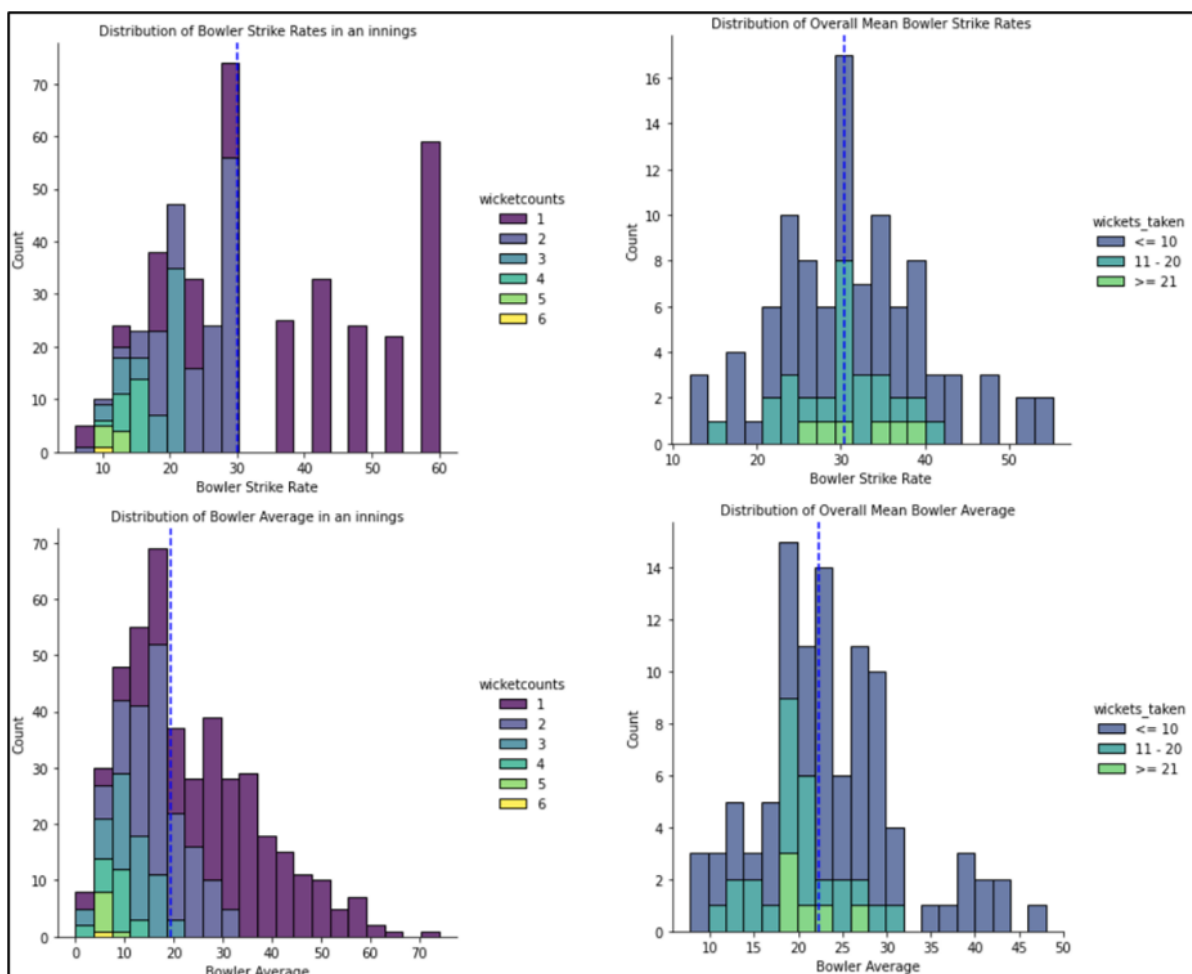


Figure 25: Bowling Strike Rate and Average in an Innings and Overall

Bowling strike rates in an innings for majority of single wicket takers ranged from 35 to 60 while for bowlers with higher wicket counts the strike rate was between 10 to 30. The overall mean strike rate throughout all the matches of the tournament ranged from 20 to 40 for bowlers with 11 or more wickets. The bowling average in an innings was on the higher side of the median for bowlers with only one wicket and some bowlers with two wickets. Higher wicket takers were on the lower



side of the median. The trend was similar for overall mean bowling average throughout the tournament. (*Figure 25*)

## 7. Identifying Key Players

The extensive process of exploratory data analysis revealed useful insights on various performance parameters of the cricket players. This has helped in identifying key performing players so far in the tournament from every team.

**Southern Vipers** - The most successful team of the Rachael Heyhoe Flint Trophy had some of the best players in their squad. GL Adams scored 733 runs throughout the fifteen matches she played in two seasons that included 1 century and 6 half centuries. Her overall batting strike rate and average were 57.75 and 52.36 respectively. She also performed well in bowling, having taken 19 wickets with an economy rate of 4.28. DN Wyatt and GA Elwiss also scored considerable runs for the team. The former scored 325 runs in six matches that included 5 half-centuries. The latter played seven matches scoring 265 runs that included a century and a half-century. CM Taylor took 28 wickets and was the highest overall wicket-taker of two seasons together. She also had a low economy rate of 3.53. TG Norris took 23 wickets and CE Dean took 19 wickets, the same as GL Adams. LK Bell with 16 wickets and PJ Scholfield with 15 wickets were other impactful bowlers.

**Western Storm** – SN Luff was the highest runs scorer for the two years combined with 756 runs with 2 centuries and 6 half-centuries. Playing thirteen matches in total, she also had stable batting strike rate of 66.04 and batting average of 68.73. Although she batted in only five games, HC Knight had scored above fifty runs 5 times totalling 381 runs and had impressive figures of batting strike rate of 92.22 and batting average of 95.35. GM Hennessy was the highest wicket-taker for Western Storm taking 20 wickets with an economy rate of 4.78. She also scored a century and a half century throughout the thirteen matches she played. FMK Morris was another high performing bowler taking 16 wickets.

**Lightning** - SJ Bryce and KE Bryce were top players with the bat having scored 597 and 494 runs respectively over thirteen innings that included one century. The former also scored 5 half-centuries and the latter scored 3. KE Bryce also took 24 wickets, the highest for Lightning and her economy rate was 3.84. 19 wickets were taken by KL Gordon with an economy rate of 3.51. LF Higham, T Graves and S Munro were other notable bowlers with 13, 12 and 10 wickets respectively.

**Northern Diamonds** – SL Kalis was the net highest run scorer with 487 runs inclusive of 5 half-centuries from fourteen innings. The second highest scorer HJ Armitage attained 373 runs from sixteen innings with 2 half-centuries. L Winfield and NR Sciver scored 1 century each. BA Langston with 25 wickets and 3.81 economy rate was the best performing bowler. She was also the second highest wicket-taker of the two years combined. KA Levick was also impactful with 23 wickets and

3.67 economy rate. JL Gunn, LCN Smith and KH Brunt had 17,16 and 15 wickets respectively in their name.

**Sunrisers** – Most batters were moderate run scorers. The highest net runs scored was 357 by CL Griffith in eleven matches. She scored 2 half-centuries. FC Wilson was the only other player with 2 half-centuries, and she appeared in eight innings making 205 runs. Sonali Patel and KS Castle were notable performers with the ball taking 14 and 11 wickets respectively.

**Central Sparks** – E Jones scored highest runs over both seasons with 633 that included 2 centuries and 4 half-centuries in fourteen matches. The following two high run scorers were Marie Kelly with 405 and GM Davies with 362 runs. The former scored 4 and the latter scored 2 half-centuries. AE Jones played just five innings yet scored centuries twice. Her net runs total was 320 and had impressive batting strike rate of 77.83 and average of 80. Five bowlers who took ten or more wickets throughout both seasons were IECM Wong with 17 wickets, followed by EL Arlott with 14, GK Davis with 13, RA Fackrell with 11 and EA Russell with 10 wickets.

**North West Thunder** – The highest runs scoring batter was GEB Boyce with 337 from thirteen matches that included half-centuries twice. The other players were moderate run scorer. EL Lamb played in seven innings and scored century once. A Hartley took 21 wickets followed by HE Jones with 18. S Ecclestone and KL Cross had 12 and 10 wickets respectively.

**South East Stars** – The batters of South East Stars were low to moderate run scorers. Significant performance with the bat was done by SIR Dunkley who despite only playing five innings, managed to score a century and 2 half centuries. Her runs total was 298 with strike rate of 67.46 and average of 74.5. AN Davidson-Richards was the only player who scored more half-centuries than her with 3 from ten innings. NE Farrant took 18 wickets, BF Smith took 16 followed by D Gregory with 11 wickets.

## 8. Scoring Players and Teams

The next task in the project focuses on using the multiple parameters of cricket observed through exploratory analysis to assign a numerical score value to every player in the tournament that reflects their batting and bowling ability.

At first for every applicable player, the values of bowling parameters such as economy rate, bowling strike rate and bowling average were inversed because unlike wickets or batting parameters, lesser values of these bowling parameters signify better performance. Thereafter all the batting and bowling parameters were scaled by Z-score normalization for consistency. This was done because different parameters have different ranges and significance of values. Z –score normalization is method of standardising values such that the mean becomes 0 and standard deviation 1. This normalization

process is also helpful for handling outlier values (Zach, 2021). Batting score and bowling scores were calculated by assigning weightages to the scaled batting and bowling parameters.

For batting score, the parameters along with their weightages were– No. of Innings Batted: 5%, Runs: 45%, No. of Half-Centuries: 5%, No. of Centuries: 5%, Batting Strike Rate: 20%, Batting Average: 20%. Therefore, the player batting score is given by

$$\begin{aligned} \text{Player Batting Score} = & 0.05 * (\text{No. of Innings Batted}) + 0.45 * (\text{Runs}) + \\ & 0.05 * (\text{No. of Half - Centuries}) + 0.05 * (\text{No. of Centuries}) + \\ & 0.20 * (\text{Batting Strike Rate}) + 0.20 * (\text{Batting Average}) \end{aligned}$$

For bowling score, the parameters along with their weightages were– No. of Overs Bowled: 5%, Wickets: 40%, Economy Rate: 25%, Bowling Strike Rate: 15%, Bowling Average: 15%. Therefore, the player bowling score is given by

$$\begin{aligned} \text{Player Bowling Score} = & 0.05 * (\text{No. of Overs Bowled}) + 0.40 * (\text{Wickets}) + \\ & 0.25 * (\text{Economy Rate}) + 0.15 * (\text{Bowling Strike Rate}) + \\ & 0.15 * (\text{Bowling Average}) \end{aligned}$$

Finally, the batting and bowling scores were scaled using the min-max normalization method to bring the scores within the 0 to 1 range. The minimum score got assigned 0 and the maximum score became 1 and the scores in between got decimal values between 0 and 1 accordingly.

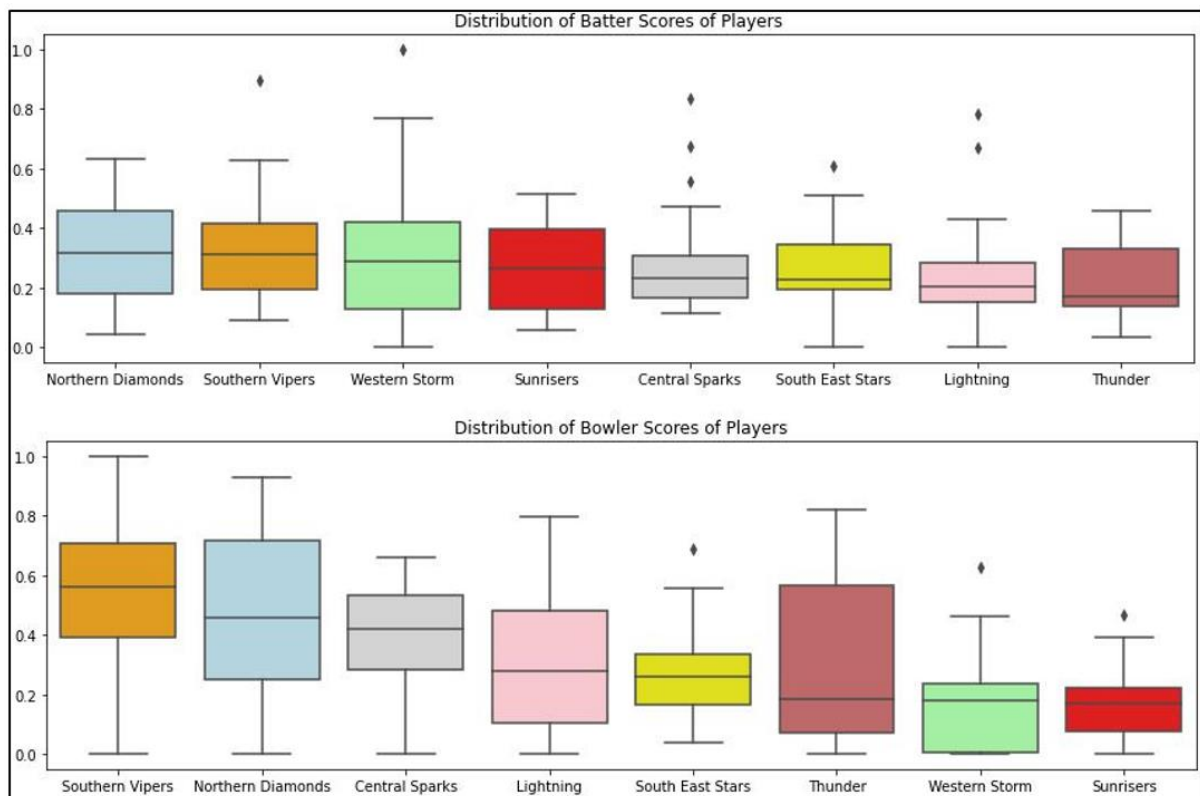


Figure 26: Distribution of Player Batting and Bowling Scores

Most batting scores of players ranged between 0.1 and 0.6 for all the teams. Few teams had players with batting scores above 0.6. Northern Diamonds, Southern Vipers and Western Storm had most players with higher batting scores that have been depicted in the upper whiskers of their boxplots as shown *Figure 26*. The best performing batters of Southern Vipers, Western Storm, Central Sparks,

South East Stars and Lightning had distinct top batting scores and they were depicted as outliers in their respective teams' box plots. The player bowling scores of the teams showed more variability. The scores mostly ranged from 0 to 0.8. The Southern Vipers had the highest median player bowling score close to 0.6 while the Sunrisers had the lowest median of around 0.2.

Finally, the next part of scoring was to obtain the team batting and team bowling scores of the playing teams of each match played in the tournament. For every match both teams had their squad of eleven players that included designated batters, allrounders and bowlers. Over the two years of the tournament a total of 55 matches were played and most teams made small changes in their playing eleven throughout their matches. The Team Batting Score of any team in a particular match was taken by averaging all the player batting scores of the batters and allrounders in that team's playing eleven. Likewise, the Team Bowling Score of any team was obtained by taking mean of the bowling scores of bowlers and allrounders among those eleven players of that match. For all the 55 matches, both playing teams had their respective team batting and bowling scores. The Team Batting Scores in most of the matches ranged from 0.25 to 0.55 and had less variability among the teams. The bowling scores ranged from 0.1 to 0.7. The Southern Vipers had the most high-valued Team Bowling Scores while the Sunrisers had the set of lowest values. Generally, teams with bowling scores greater than 0.55 tended to win while those with bowling scores less than 0.25 tended to lose the match. (Figure 27)

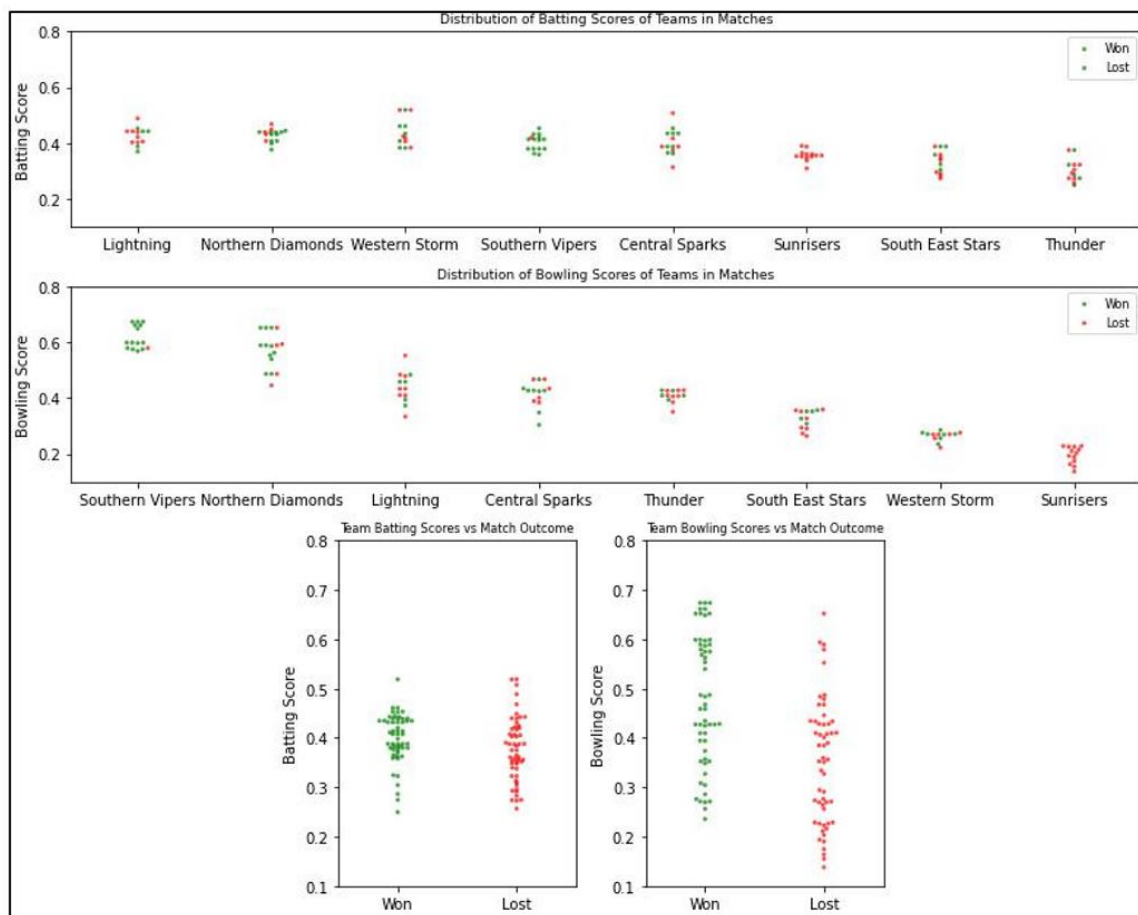


Figure 27: Distribution of Team Batting and Bowling Scores

The Team Batting Score and Team Bowling Score were part of the features used for predictive modelling as detailed in the following section.


## 9. Predictive Modelling

The final segment of this project focused on finding ways to predict the winner of every match and the second innings total of the team batting second while chasing the target runs. The match data of 55 matches played in the 2020 and 2021 seasons were used in training machine learning models. The performance of these models was then evaluated by using them to predict the match winners or second innings runs totals of the 16 matches that have already been played for 2022 season as on 23<sup>rd</sup> July. The predicted results were then validated with the actual outcomes to understand how close the predictions have been.

### 9.1 Predicting Match Winners

#### 9.1.1 Using Team Batting and Bowling Scores

**Feature and Target variables-** In this approach the Team Batting Scores and Team Bowling Scores of the playing teams have been used as feature variables along with the teams themselves. The winning teams of the matches were the target variables. The team names being categorical variables, were required to be converted to numerical form prior to put into modelling. This conversion was done by one-hot encoding method where each of the team names were represented by a separate column and binary value of 1 or 0 were assigned under them. 1 signified that the team was playing while 0 meant otherwise. *Figure 28* shows the transformation of the categorical column ‘Team1’ after one-hot encoding.



Team1	Team1__Central Sparks	Team1__Northern Diamonds	Team1__South East Stars	Team1__Southern Vipers	Team1__Sunrisers	Team1__Thunder	Team1__Western Storm
Sunrisers	0	0	0	0	1	0	0
Central Sparks	1	0	0	0	0	0	0
Thunder	0	0	0	0	0	1	0
South East Stars	0	0	1	0	0	0	0
Southern Vipers	0	0	0	1	0	0	0

Figure 28: One-hot encoding

**Modelling** – Prediction of match winners was a case of a classification problem. As the sample dataset was small with just 55 rows, a simple estimator like Linear SVC (Support Vector Classifier) was used. Linear SVC is a type of Support Vector Machine (SVM) estimator. It attempts to separate different points using a hyperplane that is most suitable to classify the data based on distinct features. The hyperplane has two parallel margins on the either side and they pass through the points closest to the hyperplane or support vectors. The greater the distance between the two margins, the better the hyperplane is for classification. (*Figure 29*)

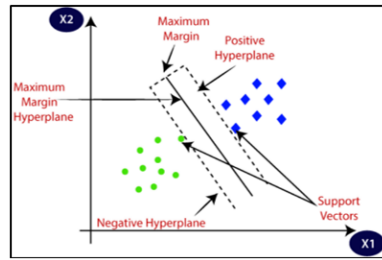


Figure 29: SVM Classification (Saini, 2021)

The input data was further split into train and test datasets, where the train dataset was proportionally larger than the test. The train dataset was used to fit the model which was then used to predict the target values for the test dataset. The purpose of the split was to equip the model to handle unseen data better and minimize chances of overfitting with the input dataset. Linear SVC models with different ratios of train-test split were trained and overall model accuracies of the test predictions were observed. For each model, cross-validation technique was also applied where that model was trained on different split versions of train data and evaluated on corresponding different split versions of test data. The average accuracy for different versions were observed. *Table 1* outlines the test accuracies found by three best versions of models. The confusion matrices in *Figure 30* shows model predicted winners of test dataset to the actual winning teams.

Model	Train-Test Split	Model Accuracy (Test Predictions)	Cross-Validated Accuracy (Test Predictions)
Linear SVC	75%-25%	64.29%	76.24%
	70%-30%	52.94%	
	80%-20%	81.82%	

Table 1: Linear SVC model test accuracies (predicting winner by scores)

The model trained with 80%-20% train-test split turned out to be the most accurate for predicting winners for test data.

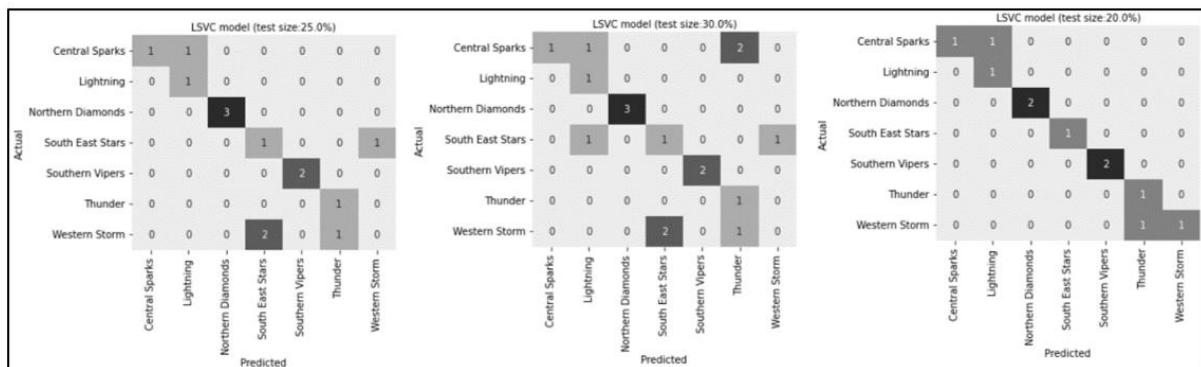


Figure 30: Confusion Matrices of Linear SVC models (predicting winner by scores)

**Validation** - The three versions of Linear SVC models were finally used to predict the winners of the 16 matches for the 2022 session and the predictions were then compared with the actual results. It was observed that the model version with 70%-30% train-test split got 81% of the winners right identifying 13 out of 16 winners correctly. The model with 75%-25% train-test split predicted 12 winners correctly while the third version predicted 11 correctly. (*Figure 31*)

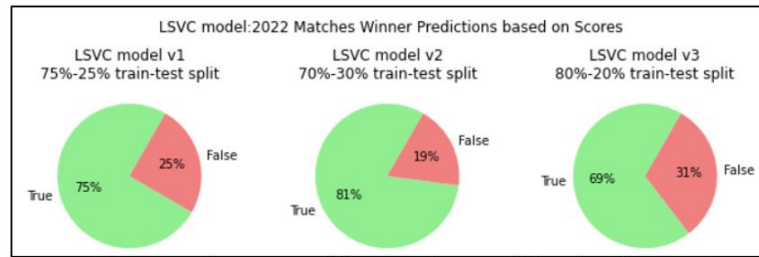


Figure 31: Linear SVC models: 2022 Winner Predictions based on Scores

**Bootstrap sampling approach-** In order to compensate for the less amount of data, the 55 rows of the input dataset was randomly sampled multiple times with repetitions to increase the row numbers. The Linear SVC model was then trained with the whole dataset without any train-test split. The trained model was then used to predict the 2022 match winners. 75% of the predictions were found to be correct and the outcome was similar to the Linear SVC model version with 75%-25% train-test split.

All prediction outcomes of each of the models have been shown in *Table 2*.

Matches	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Team1	Central Sparks	South East Stars	Western Storm	Northern Diamonds	Central Sparks	Northern Diamonds	Thunder	Southern Vipers	Central Sparks	South East Stars	Sunrisers	Southern Vipers	Southern Vipers	South East Stars	Lightning	Thunder
Team2	Southern Vipers	Sunrisers	Lightning	Thunder	Western Storm	Sunrisers	Lightning	South East Stars	Northern Diamonds	Thunder	Western Storm	Lightning	Sunrisers	Western Storm	Northern Diamonds	Central Sparks
Actual Winner	Southern Vipers	South East Stars	Western Storm	Northern Diamonds	Central Sparks	Northern Diamonds	Lightning	Southern Vipers	Northern Diamonds	South East Stars	Western Storm	Southern Vipers	Southern Vipers	South East Stars	Northern Diamonds	Thunder
LSVC (75-25)	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE
LSVC (70-30)	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE
LSVC (80-20)	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	FALSE	TRUE	FALSE
LSVC (bootstrapped)	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	FALSE

Table 2: Linear SVC model predictions of 2022 winners (predicting winner by scores)

### 9.1.2 Using Toss Outcome

**Feature and Target Variables** - In this approach, instead of batting and bowling scores, the toss winners and toss decisions at the beginning of the matches are used as feature variables to check how indicative they are in determining the target variables, winning teams at the end of the matches. One-hot encoding was applied to the categorical feature variables for numerical representation.

**Modelling** – Linear SVC estimator was used for modelling and different versions of the model were tried based on different train-splits. *Table 3* outlines the test accuracies found by three best versions of models. The confusion matrices in *Figure 32* shows model predicted winners to the actual winning teams.

Model	Train-Test Split	Model Accuracy (Test Predictions)	Cross-Validated Accuracy (Test Predictions)
Linear SVC	75%-25%	64.29%	71.15%
	70%-30%	41.18%	
	80%-20%	81.82%	

Table 3: Linear SVC model test accuracies (predicting winner by toss outcome)



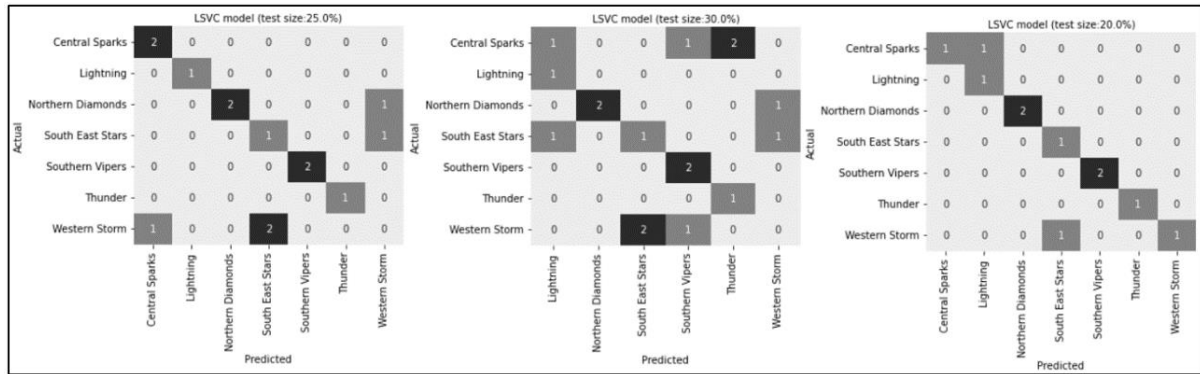


Figure 32: Confusion Matrices of Linear SVC models (predicting winner by toss outcome)

Another version of the model was also trained on bootstrapped sampled input data.

**Validation** – The three versions of the model and the bootstrapped version were used to predict the winners of the 2022 matches. The best predictions were done by the second model with 70-30% train-test split, getting 14 out of 16 or 88% of the winners correctly. The other two models with 75%-25% and 80%-20% train-test splits predicted 13 winners correctly (*Figure 33*). 12 predictions were made correctly by the model trained on bootstrapped input data.

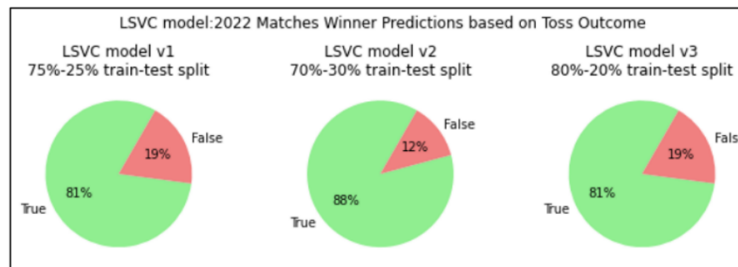


Figure 33: Linear SVC models: 2022 Winner Predictions based on Toss Outcome

All prediction outcomes of each of the models have been shown in *Table 4*.

Matches	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
<b>Team1</b>	Central Sparks	South East Stars	Western Storm	Northern Diamonds	Central Sparks	Northern Diamonds	Thunder	Southern Vipers	Central Sparks	South East Stars	Sunrisers	Southern Vipers	Southern Vipers	South East Stars	Lightning	Thunder
<b>Team2</b>	Southern Vipers	Sunrisers	Lightning	Thunder	Western Storm	Sunrisers	Lightning	South East Stars	Northern Diamonds	Thunder	Western Storm	Lightning	Sunrisers	Western Storm	Northern Diamonds	Central Sparks
<b>Actual Winner</b>	Southern Vipers	South East Stars	Western Storm	Northern Diamonds	Central Sparks	Northern Diamonds	Lightning	Southern Vipers	Northern Diamonds	South East Stars	Western Storm	Southern Vipers	Southern Vipers	South East Stars	Northern Diamonds	Thunder
<b>LSVC (75-25)</b>	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE
<b>LSVC (70-30)</b>	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE
<b>LSVC (80-20)</b>	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE
<b>LSVC (bootstrapped)</b>	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	FALSE

Table 4: Linear SVC model predictions of 2022 winners (predicting winner by toss outcome)

## 9.2 Predicting Second Innings Runs

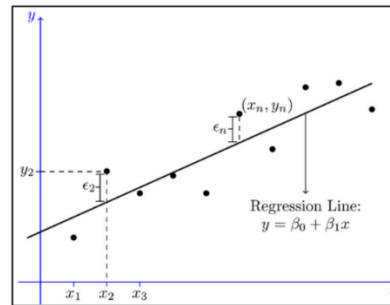
### 9.2.1 Using Team Batting and Bowling Scores

**Feature and Target variables-** The feature variables were the defending team's bowling score, the chasing team's batting score, the target runs set by the previous batting team in the first innings and the names of the playing teams. One-hot encoding was applied to the categorical features such as the



team names. The aim was to see how well the actual runs total of the second innings could be predicted from the given features.

**Modelling-** This was a case of regression problem and as the dataset is small, simple models like Linear Regression and Ridge Regression models were used. Linear Regression is the simplest machine learning algorithm where the relation between the target and the feature variables is determined by fitting a straight line through the data points, that represents a linear equation for the relationship. If most of the data points are reasonably close to the straight line then the margin of error is reduced and better outcome is achieved. (Figure 34)



Ridge Regression is a modification of the Linear Regression algorithm that has the capability to handle issues like data overfitting and multicollinearity of feature variables. The latter is especially relevant because a downside of one-hot encoding is that it some features may have the issue of multicollinearity where it becomes possible to detect value of one feature variable from another. For example, if the bowling team is ‘Lightning’ or ‘Team1\_Lightning is 1, then it is evident that there can be no other bowling teams and so the values under those columns will obviously be 0.

The input data set were split into train and test sets in different ratios for different versions of each model. Cross validation was applied on the train-test versions of the model where the splitting of the dataset was done at different points. The features and target variables of the train sets were used to fit the model. Then predictions were made on the target variables of the test set which was then evaluated with the actual values. The test accuracies found by three best versions of models as well as the RMSE (Root Mean Squared Errors) are presented in Table 5. A bootstrapped sampled version for each model was also created after getting trained on bootstrap sampled versions of the whole input data.

Model	Train-Test Split	Model Accuracy (Test Predictions)	Cross-Validated Accuracy (Test Predictions)	Error (RMSE)
Linear Regression	85%-15%	79.08%	22.78%	21.42
	80%-20%	75.74%		22.70
	90%-10%	92.16%		12.79
Ridge Regression	85%-15%	83.37%	43.31%	19.10
	80%-20%	80.56%		20.32
	90%-10%	92.00%		12.92

Table 5: Linear Regression and Ridge Regression model test accuracies (predicting by scores)

**Validation** – All versions of the Linear Regression and Ridge Regression models were used to predict the second innings runs of the 2022 matches and the results were compared. *Figure 35* compares the predictions to the actual results.

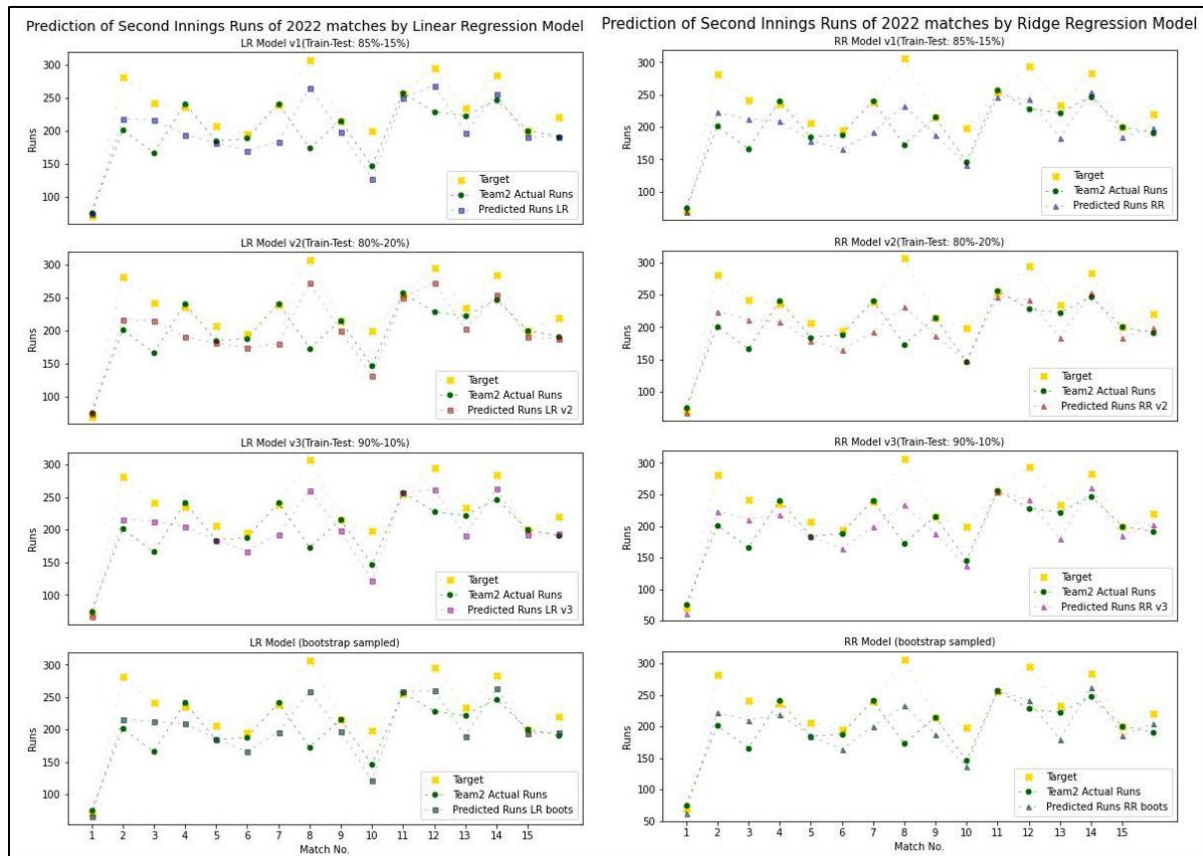


Figure 35: Linear Regression and Ridge Regression model Predictions based on Scores

In *Table 6*, the predictions that are closer to the actual are depicted in blue while the far-off points in pink. Most of the time, the predictions have been far-off. When the second innings runs exceeded the target runs, then that batting team won the match. The cells shaded red signify the team had lost while green shaded cells denote that the team had won the match. So, we can also understand the implications of the predictions as many of the predictions gave the wrong notion that the team had lost while the team had actually won and vice versa. The models' predictions were misleading in most of the situations when the batting team won the match.

Matches	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Target	70	282	242	236	207	195	239	307	215	199	255	295	234	284	200	220
Second Innings Actual Runs	75	201	166	241	184	188	241	173	215	146	257	228	222	247	200	191
LR (85-15)	74	217	216	194	181	169	183	265	198	127	250	268	197	255	190	191
LR (80-20)	74	217	215	190	181	174	180	272	199	132	250	273	202	254	190	188
LR (90-10)	67	216	213	205	184	167	193	260	198	122	257	262	191	263	192	194
LR (bootstrapped)	66	216	212	209	185	166	196	258	197	121	258	260	190	263	194	196
RR (85-15)	68	223	212	209	179	166	192	233	188	141	247	243	183	254	184	199
RR (80-20)	67	224	212	208	179	165	192	232	187	147	247	242	184	254	184	199
RR (90-10)	62	223	211	218	183	164	199	234	188	137	255	242	180	261	185	203
RR (bootstrapped)	62	222	210	219	185	164	200	233	188	136	257	241	180	262	186	205

Table 6: Linear Regression and Ridge Regression Second Innings Runs Predictions (based on Scores)

## 9.2.2 Using Information of First 10 overs

**Feature and Target variables** – In this approach instead of scores, the runs scored and wickets lost by the batting team in the first 10 overs of the second innings were used as feature variables along with the target runs and team names.

**Modelling**- Multiple train-test split versions of the Linear Regression and Ridge Regression models were explored and evaluated with the test values. *Table 7* shows the results of the three best performing models of each type. Bootstrapped sampled versions were also trained.

Model	Train-Test Split	Model Accuracy (Test Predictions)	Cross-Validated Accuracy (Test Predictions)	Error (RMSE)
Linear Regression	75%-25%	49.51%	27.45%	33.73
	80%-20%	24.84%		34.41
	90%-10%	45.54%		37.22
Ridge Regression	75%-25%	50.32%	48.30%	33.45
	80%-20%	30.02%		33.20
	90%-10%	49.05%		36.00

Table 7: Linear Regression and Ridge Regression model test accuracies (predicting by first 10 overs)

**Validation** –*Figure 36* compares the predictions for 2022 second innings runs with the actual runs scored by the second batting teams. From *Table 8*, it was noticed that some predictions were closer to the actual values than the earlier models based on scores. There is however no major difference in the performances of the Linear and Ridge Regression models regarding predicting second innings runs.

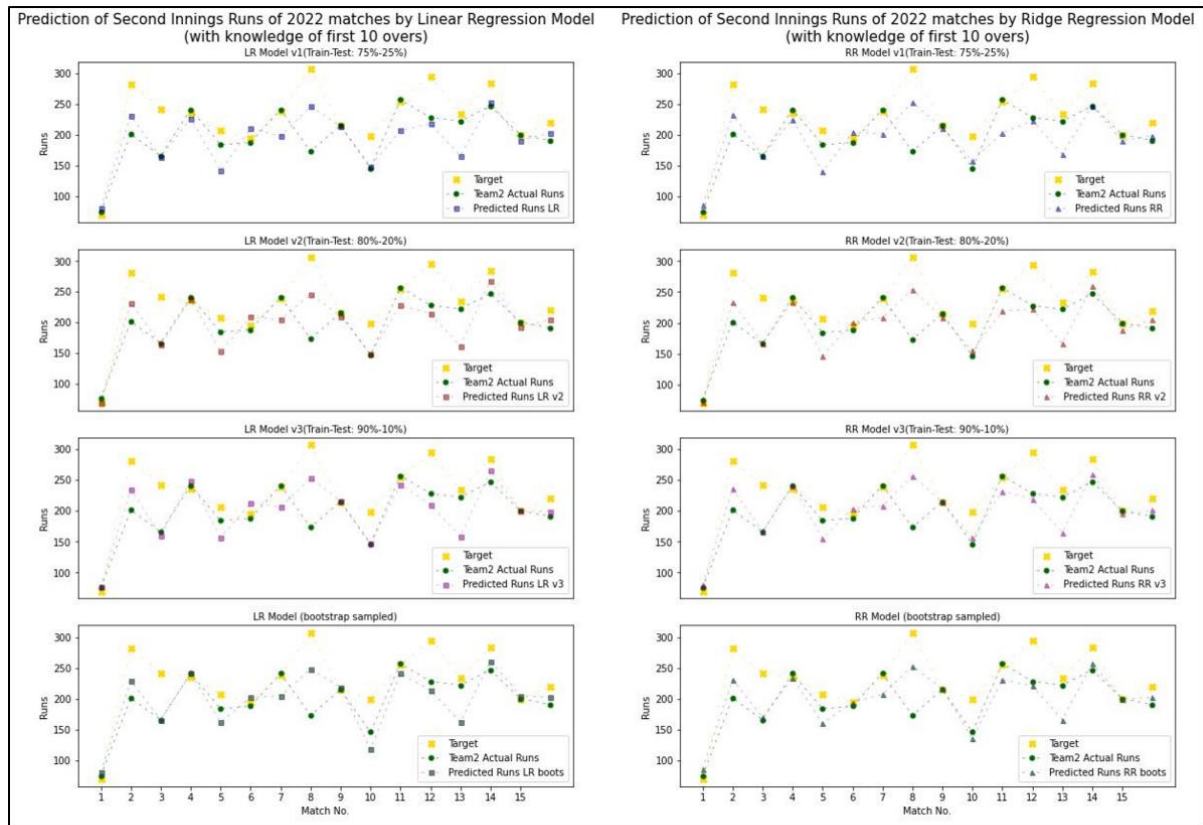


Figure 36: Linear Regression and Ridge Regression model Predictions based on First 10 overs

Matches	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Target	70	282	242	236	207	195	239	307	215	199	255	295	234	284	200	220
Second Innings Actual Runs	75	201	166	241	184	188	241	173	215	146	257	228	222	247	200	191
LR (75-25)	82	231	164	227	142	210	198	246	214	149	207	218	166	253	191	203
LR (80-20)	68	232	163	238	152	209	204	246	210	148	228	214	161	268	192	204
LR (90-10)	77	235	160	249	157	212	206	253	216	147	243	210	158	265	200	199
LR (bootstrapped)	81	229	166	241	162	202	205	248	219	118	241	214	162	261	205	202
RR (75-25)	86	233	166	225	140	204	202	253	211	158	203	223	169	247	190	199
RR (80-20)	71	233	167	234	147	201	208	254	209	155	219	222	166	260	189	205
RR (90-10)	80	236	166	241	155	204	208	257	214	156	232	219	165	260	195	201
RR (bootstrapped)	86	231	170	234	160	194	207	253	216	135	231	221	166	257	200	202

Table 8: Linear Regression and Ridge Regression Second Innings Runs Predictions (based on first 10 overs)

### 9.3 Discussion

The outcomes of the classification models were more satisfactory than the regression models. The regression models fared slightly better in the second case when the first 10 over runs and wickets were used as features instead of the batting and bowling scores. The batting and bowling scores especially the former lacked sufficient variability to be significantly influential in predicting second innings runs by regression while they were more effective in classification for predicting the winner team. The scoring method starting from the assigning weightages to player parameters to calculating team scores was applied in this project for modelling convenience. The basis of the scores were player performances that were related only to the Rachael Heyhoe Flint Trophy. There could be likelihood that a good performing player in general, might not have performed as well in the tournament and so her scores were low and that impacted the overall team scores as well. Also, some players may have scored low due to less experience of playing in the tournament.

For both classification and regression, it was observed that better predictions on the validation data or the 2022 match data were made by those models that had low test accuracy. This implied that those particular models were much better generalised or equipped to handle new and unknown data. The most likely reason is the fact that since the source input data contained only 55 rows for 55 matches played over the years 2020 and 2021. And due to small size, the models were prone to overfitting when they were trained with train dataset. Low cross validation accuracies especially for regression models were also indicative of the over fitting issues. When the test size or the unknown component increased then the test accuracy decreased yet the models developed comparatively better in terms of generalisation. The bootstrap sampling did not cause any significant improvement in the modelling process. While it did add volume, the data did not gain any variability and same rows from the 55 original data were sampled and repeated.

## 10. Conclusion

In this project a thorough analysis was conducted on multiple cricket parameters in relation to batting and bowling. The extensive exploratory analysis and visual representation helped in identifying how well the teams fared and the reasons for their success and failure. The general batting strengths of most teams were synonymous however few top run scorers and partnership contributors were influential in elevating the batting performances of their teams. Bowling performances were comparatively more indicative of a team's strength. High wicket-taking bowlers with low economy rates were assets of successful teams like Southern Vipers and Northern Diamonds.

The outcomes of the predictive modelling were mixed with the classification models providing decent estimations of the winning team but the results of the regression models regarding prediction of the second innings run total by the batting team were inconsistent. Much of it was due to small dataset size. The Rachael Heyhoe Flint Trophy is a recent tournament that started only in 2020. In the coming years, more matches will be played, more players will join or switch teams and more variety of results will be observed. It is expected that with more data in the future, there will be more avenues to explore more sophisticated methods in terms of statistical predictions that have been avoided in this project due to over-engineering concerns on a small dataset. However, the fact remains that limited overs cricket matches are rarely predictable even intuitively and there is no perfect algorithm for predicting match outcomes before the match or an innings is being played. The match result is never known until the last ball of the final over is bowled.

## 11. References

- Abhigyan (2020). *Understanding The Linear Regression!!!!* [online] Analytics Vidhya. Available at: <https://medium.com/analytics-vidhya/understanding-the-linear-regression-808c1f6941c0> [Accessed 16 Aug. 2022].
- Adhikari, A., Saraf, R. and Parma, R. (2017). *Bowling strategy building in limited over cricket match: An application of statistics*. In: MathSport International 2017 Conference Proceedings.
- Bailey M., and Clarke S.R. (2006). *Predicting the match outcome in one-day international cricket matches, while the game is in progress*. Journal of sports science & medicine, 5, 480-7
- Bandulasiri, A., Brown, T. and Wickramasinghe, I. (2016). *FACTORS AFFECTING THE RESULT OF MATCHES IN THE ONE DAY FORMAT OF CRICKET*. Operations Research & Decisions, 26(4), pp.21–33.
- BBC (2013). *Women's cricket grows in popularity*. [online] Available at: <https://www.bbc.co.uk/sport/cricket/24532779> [Accessed 15 Jun. 2022].
- Dawson, P., Morely, B., Paton, D. and Thomas, D. (2009) *To bat or not to bat: An examination of match outcomes in day-night limited overs cricket*. Journal of the Operational Research Society, vol. 60, no. 12, pp. 1786-1793.
- Dubey, P.K., Suri, H., Gupta, S. (2021). *Naïve Bayes Algorithm Based Match Winner Prediction Model for T20 Cricket*. In: Dash, S.S., Das, S., Panigrahi, B.K. (eds) Intelligent Computing and Applications. Advances in Intelligent Systems and Computing, vol 1172. Springer, Singapore. [https://doi.org/10.1007/978-981-15-5566-4\\_38](https://doi.org/10.1007/978-981-15-5566-4_38)
- England and Wales Cricket Board (2019). *Women's T20 cricket to be included in the Birmingham 2022 Commonwealth Games*. [online] <https://www.ecb.co.uk/>. Available at: <https://www.ecb.co.uk/news/1310567/womens-t20-cricket-to-be-included-in-the-birmingham-2022-commonwealth-games> [Accessed 17 Jun. 2022].
- ICC Media Release (2022). *The rising profile of women's cricket*. [online] [www.icc-cricket.com](http://www.icc-cricket.com). Available at: <https://www.icc-cricket.com/media-releases/2494728> [Accessed 16 Jun. 2022].
- Kampakis, S. and Thomas, W. (2015). *Using Machine Learning to Predict the Outcome of English County twenty over Cricket Matches*. [online] Available at: <https://arxiv.org/ftp/arxiv/papers/1511/1511.05837.pdf> [Accessed 25 Jun. 2022].

Lewis, A.J. (2005). *Towards fairer measures of player performance in one-day cricket*. Journal of the Operational Research Society, 56(7), pp.804–815. doi:10.1057/palgrave.jors.2601876.

Marylebone Cricket Club (2017). *Run out Law / MCC*. [online] www.lords.org. Available at: <https://www.lords.org/mcc/the-laws-of-cricket/run-out> [Accessed 31 Jul. 2022].

Mathews, M. (2021). *The Hundred's opening match sets UK viewing record for women's cricket*. [online] <https://www.skysports.com/>. Available at: <https://www.skysports.com/cricket/news/15821/12361385/the-hundreds-opening-match-sets-uk-viewing-record-for-womens-cricket> [Accessed 17 Jun. 2022].

Mukherjee, S. (2013). *Complex network analysis in cricket : Community structure, player's role and performance index*. Ithaca: Cornell University Library [online] arXiv.org. Available at: <https://www.proquest.com/working-papers/complex-network-analysis-cricket-community/docview/2081775415/se-2> [Accessed 20 Jun. 2022].

Nicholson, R. (2021). *Flint [née Heyhoe], Rachael, Baroness Heyhoe Flint [known as Rachael Heyhoe Flint] (1939–2017), cricketer and hockey player*. Oxford Dictionary of National Biography, pp. Oxford Dictionary of National Biography, 2021–01-14.

Passi, K. and Pandey, N. (2018). *Increased Prediction Accuracy in the Game of Cricket Using Machine Learning*. International Journal of Data Mining & Knowledge Management Process, 8(2), pp.19–36. doi:10.5121/ijdkp.2018.8203.

Reade, J., Singleton, C. and Jewell, S. (2020) *It's Just Not Cricket: The Uncontested Toss and the Gentleman's Game*. Forthcoming chapter. *Advances in Sports Economics*. Agenda Publishing, Ed: R. Butler. [online] <https://www.researchgate.net/> Available at: [https://www.researchgate.net/publication/341619082\\_It's\\_Just\\_Not\\_Cricket\\_The\\_Uncontested\\_Toss\\_and\\_the\\_Gentleman's\\_Game](https://www.researchgate.net/publication/341619082_It's_Just_Not_Cricket_The_Uncontested_Toss_and_the_Gentleman's_Game) [Accessed 18 Jun. 2022].

Saikia, H. (2020) *Quantifying the Current Form of Cricket Teams and Predicting the Match Winner*, Management and Labour Studies, 45(2), pp. 151–158. doi: 10.1177/0258042X20912603.

Saini, A. (2021). *Support Vector Machine (SVM): A Complete guide for beginners*. [online] Analytics Vidhya. Available at: <https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beginners/#:~:text=Margin%3A%20it%20is%20the%20distance> [Accessed 15 Aug. 2022].

Shah, P. and Shah, M. (2015). *Predicting ODI Cricket Result*. Journal of Tourism, Hospitality and Sports, Vol 5. [online] <https://www.researchgate.net>. Available at: [https://www.researchgate.net/publication/326187483\\_Predicting\\_ODI\\_Cricket\\_result](https://www.researchgate.net/publication/326187483_Predicting_ODI_Cricket_result) [Accessed 20 Jun. 2022].

Shvili, J. (2020). *The Most Popular Sports in the World*. [online] WorldAtlas. Available at: <https://www.worldatlas.com/articles/what-are-the-most-popular-sports-in-the-world.html> [Accessed 15 Jun. 2022].

Velija, P. (2015). *Women's Cricket and Global Processes: The Emergence and Development of Women's Cricket as a Global Game*. London: Palgrave Macmillan.

Velija, P., Ratna, A. and Flintoff, A. (2012). *Exclusionary power in sports organisations: The merger between the Women's Cricket Association and the England and Wales Cricket Board*. International Review for the Sociology of Sport, 49(2), pp.211–226. doi:10.1177/1012690212455962.

Veppur Sankaranarayanan, V. (2014). *Towards a time-lapse prediction system for cricket matches*. [online] open.library.ubc.ca. Available at: <https://open.library.ubc.ca/soa/cIRcle/collections/ubctheses/24/items/1.0167478> [Accessed 20 Jun. 2022].

Wickramasinghe, I. (2020a). *Naive Bayes approach to predict the winner of an ODI cricket game*. Journal of Sports Analytics, vol. 6(2), pp.75–84. doi:10.3233/jsa-200436.

Wickramasinghe, I. (2020b). *Classification of All-Rounders in the Game of ODI Cricket: Machine Learning Approach*. ATHENS JOURNAL OF SPORTS, 7(1), pp.21–34. doi:10.30958/ajspo.7-1-2.

Wikipedia. (2022a). *Rachael Heyhoe Flint Trophy*. [online] Available at: [https://en.wikipedia.org/wiki/Rachael\\_Heyhoe\\_Flint\\_Trophy](https://en.wikipedia.org/wiki/Rachael_Heyhoe_Flint_Trophy) [Accessed 17 Jun. 2022].

Wikipedia. (2022b). *Strike rate*. [online] Available at: [https://en.wikipedia.org/wiki/Strike\\_rate](https://en.wikipedia.org/wiki/Strike_rate). [Accessed 17 Jul. 2022].

Wikipedia. (2022c). *Batting average (cricket)*. [online] Available at: [https://en.wikipedia.org/wiki/Batting\\_average\\_\(cricket\)](https://en.wikipedia.org/wiki/Batting_average_(cricket)) [Accessed 17 Jul. 2022].

Wikipedia. (2022d). *Economy rate*. [online] Available at: [https://en.wikipedia.org/wiki/Economy\\_rate](https://en.wikipedia.org/wiki/Economy_rate) [Accessed 17 Jul. 2022]



Wikipedia. (2022e). *Strike rate*. [online] Available at:  
[https://en.wikipedia.org/wiki/Strike\\_rate#:~:text=Bowling%20strike%20rate%20is%20defined,is%20at%20taking%20wickets%20quickly](https://en.wikipedia.org/wiki/Strike_rate#:~:text=Bowling%20strike%20rate%20is%20defined,is%20at%20taking%20wickets%20quickly). [Accessed 17 Jul. 2022]

Wikipedia. (2022f). *Bowling average*. [online] Available at:  
[https://en.wikipedia.org/wiki/Bowling\\_average](https://en.wikipedia.org/wiki/Bowling_average) [Accessed 17 Jul. 2022]

Zach (2021). *Z-Score Normalization: Definition & Examples*. [online] Statology. Available at:  
<https://www.statology.org/z-score-normalization/#:~:text=Z%2Dscore%20normalization%20refers%20to>. [Accessed 25 Jul. 2022]

## 12. Appendices

### Appendix 12.1: Top 20 runs scoring batters in an innings

2020 season

BATTER	RUNS	OPPONENT (DATE)
GL Adams (Southern Vipers)	154	vs Western Storm (13/09/2020)
SJ Bryce (Lightning)	136	vs Central Sparks (19/09/2020)
E Jones (Central Sparks)	115	vs Lightning (19/09/2020)
GM Hennessy (Western Storm)	105	vs Sunrisers (19/09/2020)
NR Sciver (Northern Diamonds)	104	vs Lightning (31/08/2020)
SN Luff (Western Storm)	104	vs South East Stars (11/09/2020)
AD Carr (Sunrisers)	99	vs South East Stars (31/08/2020)
SIR Dunkley (South East Stars)	97	vs Sunrisers (31/08/2020)
AL MacDonald (Northern Diamonds)	92	vs Thunder (10/09/2020)
HC Knight (Western Storm)	91	vs South East Stars (29/08/2020)
E Jones (Central Sparks)	90	vs Thunder (05/09/2020)
GL Adams (Southern Vipers)	89	vs Sunrisers (11/09/2020)
SL Kalis (Northern Diamonds)	87	vs Central Sparks (13/09/2020)
GM Hennessy (Western Storm)	86	vs South East Stars (11/09/2020)
SN Luff (Western Storm)	85	vs Sunrisers (05/09/2020)
A Griffiths (Western Storm)	80	vs Sunrisers (19/09/2020)
GL Adams (Southern Vipers)	80	vs Northern Diamonds (27/09/2020)
C Brewer (South East Stars)	79	vs Southern Vipers (19/09/2020)
SN Luff (Western Storm)	79	vs Southern Vipers (13/09/2020)
E Jones (Central Sparks)	77	vs Northern Diamonds (13/09/2020)

Table 9: Top 20 runs scoring batters in an innings in 2020

2021 season

BATTER	RUNS	OPPONENT (DATE)
AE Jones (Central Sparks)	163	vs Western Storm (31/05/2021)
KE Bryce (Lightning)	162	vs Central Sparks (18/09/2021)
SN Luff (Western Storm)	157	vs Sunrisers (18/09/2021)
EL Lamb (Thunder)	121	vs Western Storm (29/05/2021)
AE Jones (Central Sparks)	114	vs Northern Diamonds (29/05/2021)
GA Elwiss (Southern Vipers)	112	vs Sunrisers (12/09/2021)
L Winfield (Northern Diamonds)	110	vs Central Sparks (29/05/2021)
SIR Dunkley (South East Stars)	104	vs Sunrisers (29/05/2021)
E Jones (Central Sparks)	100	vs Thunder (12/06/2021)
AN Davidson-Richards (South East Stars)	92	vs Western Storm (05/06/2021)
SIR Dunkley (South East Stars)	92	vs Western Storm (05/06/2021)
GEB Boyce (Thunder)	91	vs Sunrisers (31/05/2021)
CL Griffith (Sunrisers)	91	vs Southern Vipers (12/09/2021)
HC Knight (Western Storm)	91	vs Thunder (29/05/2021)
LA Parfitt (Western Storm)	91	vs Central Sparks (31/05/2021)
SJ Bryce (Lightning)	90	vs Central Sparks (18/09/2021)
TT Beaumont (Lightning)	89	vs Thunder (06/06/2021)
KE Bryce (Lightning)	87	vs South East Stars (10/09/2021)
BF Smith (South East Stars)	84	vs Lightning (10/09/2021)
GA Elwiss (Southern Vipers)	84	vs Northern Diamonds (18/09/2021)

Table 10: Top 20 runs scoring batters in an innings in 2021

## Appendix 12.2: Top 10 overall runs scoring batters

2020 season

BATTER	RUNS
GL Adams (Southern Vipers)	500
SJ Bryce (Lightning)	395
SN Luff (Western Storm)	339
E Jones (Central Sparks)	334
Marie Kelly (Central Sparks)	223
GM Hennessy (Western Storm)	209
SL Kalis (Northern Diamonds)	197
J Gardner (Sunrisers)	193
N Brown (Thunder)	189
ME Bouchier (Southern Vipers)	183

Table 11: Top 10 overall runs scoring batters in 2020

2021 season

BATTER	RUNS
SN Luff (Western Storm)	417
KE Bryce (Lightning)	353
E Jones (Central Sparks)	299
SL Kalis (Northern Diamonds)	290
AE Jones (Central Sparks)	282
CL Griffith (Sunrisers)	273
GA Elwiss (Southern Vipers)	265
BF Smith (South East Stars)	252
EL Lamb (Thunder)	237
GL Adams (Southern Vipers)	233

Table 12: Top 10 overall runs scoring batters in 2021

## Appendix 12.3: Top 20 runs scoring batting partnerships in an innings

2020 season

PARTNERS	RUNS	OPPONENT (DATE)
GM Hennessy & SN Luff (Western Storm)	162	vs South East Stars (11-09-2020)
GL Adams & EM McCaughan (Southern Vipers)	155	vs Western Storm (13-09-2020)
A Griffiths & GM Hennessy (Western Storm)	155	vs Sunrisers (19-09-2020)
SJ Bryce & B Ellis (Lightning)	148	vs Central Sparks (11-09-2020)
L Winfield & HJ Armitage (Northern Diamonds)	139	vs Central Sparks (29-08-2020)
ME Bouchier & GL Adams (Southern Vipers)	133	vs Western Storm (13-09-2020)
HC Knight & SN Luff (Western Storm)	130	vs Southern Vipers (31-08-2020)
SN Luff & NAJ Wraith (Western Storm)	129	vs Southern Vipers (13-09-2020)
T Graves & SJ Bryce (Lightning)	127	vs Central Sparks (19-09-2020)
Marie Kelly & E Jones (Central Sparks)	123	vs Northern Diamonds (13-09-2020)
AN Davidson-Richards & SIR Dunkley (South East Stars)	123	vs Sunrisers (31-08-2020)
TT Beaumont & SJ Bryce (Lightning)	117	vs Northern Diamonds (31-08-2020)
A Capsey & A Cranstone (South East Stars)	114	vs Sunrisers (13-09-2020)
E Jones & GM Davies (Central Sparks)	114	vs Thunder (05-09-2020)
GL Adams & DN Wyatt (Southern Vipers)	111	vs Sunrisers (29-08-2020)
GL Adams & DN Wyatt (Southern Vipers)	105	vs Western Storm (31-08-2020)
E Jones & GM Davies (Central Sparks)	104	vs Lightning (19-09-2020)
GL Adams & EM McCaughan (Southern Vipers)	100	vs Northern Diamonds (27-09-2020)
MK Villiers & ND Dattani (Sunrisers)	92	vs Southern Vipers (29-08-2020)
SL Kalis & RHM Hopkins (Northern Diamonds)	85	vs Central Sparks (13-09-2020)

Table 13: Top 20 runs scoring batting partnerships in an innings in 2020

2021 season

<b>PARTNERS</b>	<b>RUNS</b>	<b>OPPONENT (DATE)</b>
KE Bryce & SJ Bryce (Lightning)	207	vs Central Sparks (18-09-2021)
P Cleary & EL Lamb (Thunder)	177	vs Western Storm (29-05-2021)
AN Davidson-Richards & SIR Dunkley (South East Stars)	154	vs Western Storm (05-06-2021)
Marie Kelly & E Jones (Central Sparks)	119	vs Lightning (18-09-2021)
SN Luff & A Griffiths (Western Storm)	119	vs Sunrisers (18-09-2021)
A Campbell & SL Kalis (Northern Diamonds)	116	vs Central Sparks (22-09-2021)
AJ Macleod & CL Griffith (Sunrisers)	112	vs Southern Vipers (12-09-2021)
SL Kalis & L Winfield (Northern Diamonds)	112	vs Central Sparks (29-05-2021)
BF Smith & AN Davidson-Richards (South East Stars)	107	vs Sunrisers (29-05-2021)
BF Smith & A Capsey (South East Stars)	105	vs Lightning (10-09-2021)
AE Jones & Marie Kelly (Central Sparks)	103	vs Northern Diamonds (29-05-2021)
LA Parfitt & HC Knight (Western Storm)	102	vs Central Sparks (31-05-2021)
GL Adams & GA Elwiss (Southern Vipers)	100	vs Sunrisers (12-09-2021)
KE Bryce & TT Beaumont (Lightning)	98	vs Thunder (06-06-2021)
ME Bouchier & GL Adams (Southern Vipers)	97	vs Western Storm (12-06-2021)
BA Langston & BAM Heath (Northern Diamonds)	91	vs Thunder (12-09-2021)
HC Knight & SN Luff (Western Storm)	91	vs South East Stars (05-06-2021)
CL Griffith & G Scrivens (Sunrisers)	90	vs Western Storm (18-09-2021)
JL Gunn & SL Kalis (Northern Diamonds)	90	vs South East Stars (12-06-2021)
GL Adams & GA Elwiss (Southern Vipers)	89	vs Northern Diamonds (18-09-2021)

Table 14: Top 20 runs scoring batting partnerships in an innings in 2021

#### Appendix 12.4: Top 10 overall runs scoring batting partnerships

2020 season

<b>PARTNERS</b>	<b>RUNS</b>
GL Adams & EM McCaughan (Southern Vipers)	401
ME Bouchier & GL Adams (Southern Vipers)	277
SJ Bryce & B Ellis (Lightning)	268
Marie Kelly & E Jones (Central Sparks)	266
E Jones & GM Davies (Central Sparks)	261
GL Adams & DN Wyatt (Southern Vipers)	216
L Winfield & HJ Armitage (Northern Diamonds)	179
GM Hennessy & SN Luff (Western Storm)	179
SN Luff & NAJ Wraith (Western Storm)	173
HC Knight & SN Luff (Western Storm)	156

Table 15: Top 10 overall runs scoring batting partnerships in 2020

2021 season

<b>PARTNERS</b>	<b>RUNS</b>
Marie Kelly & E Jones (Central Sparks)	281
KE Bryce & SJ Bryce (Lightning)	280
SN Luff & A Griffiths (Western Storm)	254
AJ Macleod & CL Griffith (Sunrisers)	250
CL Griffith & G Scrivens (Sunrisers)	232
BF Smith & AN Davidson-Richards (South East Stars)	223
ME Bouchier & GL Adams (Southern Vipers)	209
P Cleary & EL Lamb (Thunder)	189
GL Adams & GA Elwiss (Southern Vipers)	189
EL Lamb & GEB Boyce (Thunder)	189

Table 16: Top 10 overall runs scoring batting partnerships in 2021

## Appendix 12.5: Top bowlers with most wickets (4 or above) in an innings

2020 season

BOWLERS	WICKETS	OPPONENT (DATE)
CM Taylor (Southern Vipers)	6	vs Northern Diamonds (27-09-2020)
KH Brunt (Northern Diamonds)	5	vs Central Sparks (29-08-2020)
FMK Morris (Western Storm)	5	vs Sunrisers (05-09-2020)
KE Bryce (Lightning)	5	vs Northern Diamonds (31-08-2020)
A Hartley (Thunder)	4	vs Lightning (13-09-2020)
Sonali Patel (Sunrisers)	4	vs South East Stars (31-08-2020)
CM Taylor (Southern Vipers)	4	vs Western Storm (13-09-2020)
TG Norris (Southern Vipers)	4	vs Western Storm (31-08-2020)
LK Bell (Southern Vipers)	4	vs South East Stars (19-09-2020)
KE Bryce (Lightning)	4	vs Thunder (13-09-2020)
GM Hennessy (Western Storm)	4	vs South East Stars (29-08-2020)
EA Russell (Central Sparks)	4	vs Northern Diamonds (13-09-2020)
CK Boycott (Central Sparks)	4	vs Thunder (05-09-2020)

Table 17: Top bowlers with most wickets (4 or above) in an innings in 2020

2021 season

BOWLERS	WICKETS	OPPONENT (DATE)
NE Farrant (South East Stars)	5	vs Sunrisers (29-05-2021)
LCN Smith (Northern Diamonds)	5	vs Western Storm (10-09-2021)
EL Arlott (Central Sparks)	5	vs Southern Vipers (05-06-2021)
IECM Wong (Central Sparks)	5	vs Northern Diamonds (29-05-2021)
HE Jones (Thunder)	5	vs South East Stars (18-09-2021)
KE Bryce (Lightning)	4	vs Northern Diamonds (31-05-2021)
NE Farrant (South East Stars)	4	vs Western Storm (05-06-2021)
CM Taylor (Southern Vipers)	4	vs South East Stars (31-05-2021)
GL Adams (Southern Vipers)	4	vs Northern Diamonds (25-09-2021)
KL Gordon (Lightning)	4	vs Thunder (06-06-2021)
RA Fackrell (Central Sparks)	4	vs Western Storm (31-05-2021)
KL Gordon (Lightning)	4	vs Southern Vipers (29-05-2021)
Sonali Patel (Sunrisers)	4	vs Thunder (31-05-2021)
TG Norris (Southern Vipers)	4	vs Thunder (10-09-2021)
KA Levick (Northern Diamonds)	4	vs Thunder (12-09-2021)
GK Davis (Central Sparks)	4	vs South East Stars (12-09-2021)
T Graves (Lightning)	4	vs Sunrisers (12-06-2021)
KH Brunt (Northern Diamonds)	4	vs Lightning (31-05-2021)

Table 18: Top bowlers with most wickets (4 or above) in an innings in 2021

## Appendix 12.6: Top bowlers with most wickets (10 or above) overall

2020 season

BOWLERS	WICKETS
CM Taylor (Southern Vipers)	15
KE Bryce (Lightning)	14
TG Norris (Southern Vipers)	12
BA Langston (Northern Diamonds)	12
GM Hennessy (Western Storm)	11
FMK Morris (Western Storm)	11
KA Levick (Northern Diamonds)	11
A Hartley (Thunder)	11

Table 19: Top bowlers with most wickets (10 or above) overall in 2020

2021 season

BOWLERS	WICKETS
KL Gordon (Lightning)	16
HE Jones (Thunder)	14
IECM Wong (Central Sparks)	14
BA Langston (Northern Diamonds)	13
CM Taylor (Southern Vipers)	13
KA Levick (Northern Diamonds)	12
LCN Smith (Northern Diamonds)	12
BF Smith (South East Stars)	12
GL Adams (Southern Vipers)	12
RA Fackrell (Central Sparks)	11
TG Norris (Southern Vipers)	11
EL Arlott (Central Sparks)	11
A Hartley (Thunder)	10
CE Dean (Southern Vipers)	10
KE Bryce (Lightning)	10
KS Castle (Sunrisers)	10

Table 20: Top bowlers with most wickets (10 or above) overall in 2021

## Appendix 12.7: Python code for Linear SVC modelling

#function takes the features dataset(X), target dataset(y) and proportion of test dataset (testsize) as parameters

def LSVC\_model\_function(X,y,testsize):

    #import relevant libraries and packages

    from sklearn.model\_selection import train\_test\_split

    from sklearn.svm import LinearSVC

    from sklearn.model\_selection import cross\_val\_score

    from sklearn.metrics import confusion\_matrix

    #perform train test split based on specified 'testsize'

    X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=testsize, random\_state=42)

    #train the model

    model=LinearSVC(C=1.0)

    model.fit(X\_train,y\_train)

    #make predictions for test data

    y\_pred=model.predict(X\_test)

    #print accuracies and show the confusion matrix

    print("Test Accuracy: ", '%.2f'%(model.score(X\_test, y\_test)\*100)+"%")

    print(f"Cross-Validated Accuracy:{np.mean(cross\_val\_score(model, X, y, cv=4))\*100:.2f}%")

    print("Confusion Matrix: ")

    plt.title("LSVC model (test size:" + str(testsize\*100) + "%)",fontsize=10)

    cm = pd.crosstab(y\_test,y\_pred,rownames=["Actual"],colnames=["Predicted"])

    sns.heatmap(cm, annot=True, cmap=sns.color\_palette("Greys"))

    return model

#bootstrap sampled version

#function takes the features dataset(X), target dataset(y) and number of times to be sampled (n\_rep) as parameters

def LSVC\_model\_function\_bootstrapped(X,y,n\_rep):

    #import relevant libraries and packages

    from sklearn.svm import LinearSVC

    #combine X and y to replicate the original dataframe

    df = pd.concat([X, y], axis=1)

```

for i in range(n_rep):
    #sample the dataframe with repetitions of rows allowed and create new larger dataframe
    df_new = df.sample(n=len(df)+i, replace=True, random_state=42+i)

    #split the new dataframe into X_new and y_new for features and targets respectively
    X_new = df_new.drop('Winner',axis=1)
    y_new = df_new['Winner']

    #train the model with the new larger dataset
    model = LinearSVC(C=1.0)
    model.fit(X_new,y_new)

return model

```

## Appendix 12.8: Python code for Linear Regression modelling

#function takes the features dataset(X), target dataset(y) and proportion of test dataset (testsize) as parameters

```
def LinReg_model_function(X,y,testsize):
```

```
    #import relevant libraries and packages
```

```
    from sklearn.model_selection import train_test_split
```

```
    from sklearn import linear_model
```

```
    from sklearn.model_selection import cross_val_score
```

```
    from sklearn.metrics import r2_score, mean_squared_error
```

```
    #perform train test split based on specified 'testsize'
```

```
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=testsize, random_state=42)
```

```
    #train the model
```

```
    model=linear_model.LinearRegression()
```

```
    model.fit(X_train,y_train)
```

```
    #make predictions for test data
```

```
    y_pred = model.predict(X_test).astype(int)
```

```
    #print accuracies and errors
```

```
    print("Test Accuracy: ", '%.2f'%(r2_score(y_test, y_pred)*100))
```

```
    print(f"Cross-Validated Accuracy:{np.mean(cross_val_score(model, X, y, cv=3))*100:.2f}%")
```

```
    print("RMSE:%.2f" %np.sqrt(mean_squared_error(y_test, y_pred)))
```

```
    return model
```

#bootstrap sampled version

#function takes the features dataset(X), target dataset(y) and number of times to be sampled (n\_rep) as parameters

```
def LinReg_model_function_boots(X,y,n_rep):
```

```
    #import relevant libraries and packages
```

```
    from sklearn import linear_model
```

```
    #combine X and y to replicate the original dataframe
```

```
    df = pd.concat([X, y], axis=1)
```

```
    for i in range(n_rep):
```

```
        #sample the dataframe with repetitions of rows allowed and create new larger dataframe
```

```
        df_new = df.sample(n=len(df)+i, replace=True, random_state=42+i)
```

```
        #split the new dataframe into X_new and y_new for features and targets respectively
```

```
        X_new = df_new.drop('Team2_total_runs',axis=1)
```

```
        y_new = df_new['Team2_total_runs']
```

```
        #train the model with the new larger dataset
```

```
        model=linear_model.LinearRegression()
```

```
        model.fit(X_new,y_new)
```

```
    return model
```

## Appendix 12.9: Python code for Ridge Regression modelling

```
#function takes the features dataset(X), target dataset(y) and proportion of test dataset (testsize) as parameters
def LinReg_model_function(X,y,testsize):
    #import relevant libraries and packages
    from sklearn.model_selection import train_test_split
    from sklearn import linear_model
    from sklearn.model_selection import cross_val_score
    from sklearn.metrics import r2_score, mean_squared_error

    #perform train test split based on specified 'testsize'
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=testsize, random_state=42)

    #train the model
    model=linear_model.Ridge(alpha=1.0)
    model.fit(X_train,y_train)

    #make predictions for test data
    y_pred = model.predict(X_test).astype(int)

    #print accuracies and errors
    print("Test Accuracy: ", '%.2f'%(r2_score(y_test, y_pred)*100))
    print(f"Cross-Validated Accuracy:{np.mean(cross_val_score(model, X, y, cv=3))*100:.2f}%")
    print("RMSE:%.2f" %np.sqrt(mean_squared_error(y_test, y_pred)))

    return model

#bootstrap sampled version
#function takes the features dataset(X), target dataset(y) and number of times to be sampled (n_rep) as parameters
def LinReg_model_function_boots(X,y,n_rep):
    #import relevant libraries and packages
    from sklearn import linear_model

    #combine X and y to replicate the original dataframe
    df = pd.concat([X, y], axis=1)

    for i in range(n_rep):
        #sample the dataframe with repetitions of rows allowed and create new larger dataframe
        df_new = df.sample(n=len(df)+i, replace=True, random_state=42+i)
        #split the new dataframe into X_new and y_new for features and targets respectively
        X_new = df_new.drop('Team2_total_runs',axis=1)
        y_new = df_new['Team2_total_runs']

        #train the model with the new larger dataset
        model=linear_model.Ridge(alpha=1.0)
        model.fit(X_new,y_new)

    return model
```