Cricket Data scraping and analysis for robust data-driven decisions

Laiba Nadeem

UNIVERSITI TEKNOLOGI MALAYSIA

Table of Content

Chapter 4

4.1 Introduction

The results obtained from the first part of the cricket data analysis are reported in this chapter. It briefly describes how to conduct exploratory data analysis (EDA) to identify important statistics. Next, the data cleaning step in terms of missing values handling in the dataset and dataset cleaning, in general, is explained. Exploratory analysis is followed, and computed KPIs and insights from the batting and bowling data are presented. The chapter also assesses the models and algorithms employed for the player and teams' performance analysis.

4.2 Data Preprocessing:

Data cleaning was eventually the first step to maintaining the dataset's quality. This concerned gaps in and duplications of columns and formatting problems. For example, special characters in player names were removed, and leading or trailing spaces were stripped using the code:

4.2.1 Data Collection:

Four different types of data were collected for cricket data analysis. Batsman data, Bowler's data, Team matches, and matches played on different grounds. All these four aspects influence cricket match,

The data was collected from Bright Data, which utilized its data scrapping tool to scrap the cricket data from ESPN Cricinfo. The data was initially scrapped in JSON files and then transformed into CSV files using pandas. The conversion was made to make it easier to work in Power BI.

4.2.2 Bowling Data Processing:

	match	bowlingTeam	bowlerName	overs	maiden	runs	wickets	economy	0s	4s	6s	wides	noBalls	match_id
0	Namibia Vs Sri Lanka	Sri Lanka	Maheesh Theekshana	4.0	0	23	1	5.75	7	0	0	2	0	T20I # 1823
1	Namibia Vs Sri Lanka	Sri Lanka	Dushmantha Chameera	4.0	0	39	1	9.75	6	3	1	2	0	T20I # 1823
2	Namibia Vs Sri Lanka	Sri Lanka	Pramod Madushan	4.0	0	37	2	9.25	6	3	1	0	0	T20I # 1823
3	Namibia Vs Sri Lanka	Sri Lanka	Chamika Karunaratne	4.0	0	36	1	9.00	7	3	1	1	0	T20I # 1823
4	Namibia Vs Sri Lanka	Sri Lanka	Wanindu Hasaranga de Silva	4.0	0	27	1	6.75	8	1	1	0	0	T20I # 1823

Figure 4.2.2

Quantitative data that relates to bowlers were pulled out including overs bowled, the number of runs given, wickets claimed, and economy rates.
Every bowling performance was matched to the match it belongs to using a match_id from a match_id dictionary.

4.2.3 Batting Data Processing:

	match	teamInnings	battingPos	batsmanName	runs	balls	4s	6s	SR	out/not_out	match_id
0	Namibia Vs Sri Lanka	Namibia	1	Michael van Lingen	3	6	0	0	50.00	out	T20I # 1823
1	Namibia Vs Sri Lanka	Namibia	2	Divan la Cock	9	9	1	0	100.00	out	T20I # 1823
2	Namibia Vs Sri Lanka	Namibia	3	Jan Nicol Loftie-Eaton	20	12	1	2	166.66	out	T20I # 1823
3	Namibia Vs Sri Lanka	Namibia	4	Stephan Baard	26	24	2	0	108.33	out	T20I # 1823
4	Namibia Vs Sri Lanka	Namibia	5	Gerhard Erasmus(c)	20	24	0	0	83.33	out	T20I # 1823

Figure 4.2.3

Batting data initially consisted of the following columns, which gave the player's name, the runs it scored in a particular match, and other factors.

4.2.4 Players Data:

	name	team	image	battingStyle	bowlingStyle	playingRole	description
0	Najmul Hossain Shanto	Bangladesh	NaN	Left hand Bat	Right arm Offbreak	Top order Batter	Nazmul Hossain Shanto emerged from an unusual ...
1	Soumya Sarkar	Bangladesh	NaN	Left hand Bat	Right arm Medium fast	Middle order Batter	A rarity among Bangladesh allrounders, top-ord...
2	Litton Das	Bangladesh	NaN	Right hand Bat	NaN	Wicketkeeper Batter	Liton Das is the first wicketkeeper-batsman in...
3	Shakib Al Hasan(c)	Bangladesh	NaN	Left hand Bat	Slow Left arm Orthodox	Allrounder	When the annals of Bangladesh cricket are sift...
4	Afif Hossain	Bangladesh	NaN	Left hand Bat	Right arm Offbreak	Allrounder	Bangladesh left-hander Afif Hossain made his T...

Figure 4.2.4

Players' data consists of data regarding each player, the team they belong to, their balling order, and their batting order and style.

4.2.5 Match Summary Data:

	team1	team2	winner	margin	ground	matchDate	match_id
0	Namibia	Sri Lanka	Namibia	55 runs	Geelong	Oct 16, 2022	T20I # 1823
1	Netherlands	U.A.E.	Netherlands	3 wickets	Geelong	Oct 16, 2022	T20I # 1825
2	Scotland	West Indies	Scotland	42 runs	Hobart	Oct 17, 2022	T20I # 1826
3	Ireland	Zimbabwe	Zimbabwe	31 runs	Hobart	Oct 17, 2022	T20I # 1828
4	Namibia	Netherlands	Netherlands	5 wickets	Geelong	Oct 18, 2022	T20I # 1830

Figure 4.2.5

Match summary data contains a summary of different matches.

4.2.6 Data Export:

All the JSON data was converted into their respective CV files and exported. Such steps helped guarantee that the data is fit for utilization in other forms of analysis or for loading to other tools.

4.3 Exploratory Data Analysis (EDA):

Exploratory analysis was first done to identify the data type being dealt with. These datasets involved match summaries, players' profiles, and comprehensive Batting and Bowling statistics. The datasets and their key patterns and trends are summarized below:

4.3.1 Match Outcomes:

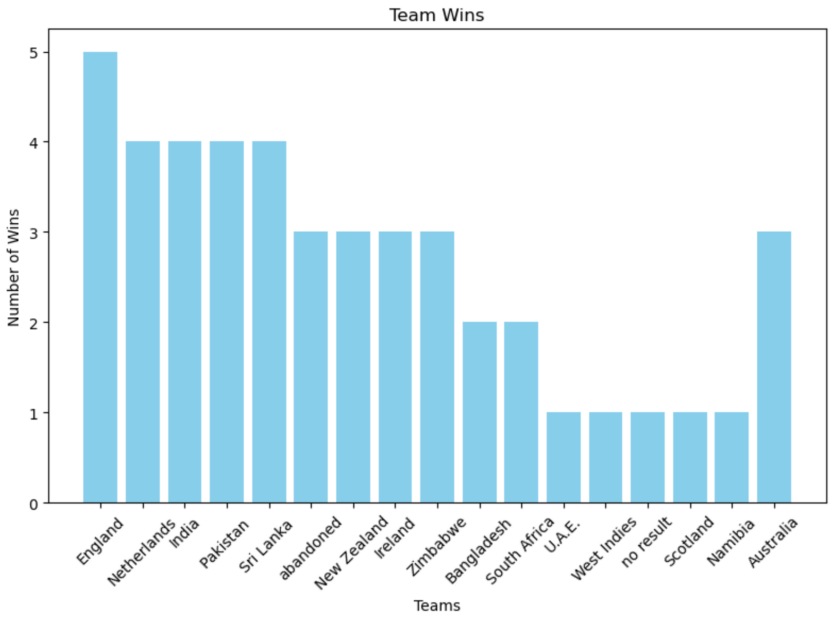

Figure 4.2.8 (a)

Figure 4.2.8(a) represents the dataset of different T20 matches between the teams and their results. From different analyzed match outcomes, it was seen that England is one of the most dominant teams in the dataset, with the most wins to their credit. Netherlands, Pakistan, Sri Lanka, and India picked four wins apiece. These results indicate closely associated competitive performances of two dominant teams in the tournament.

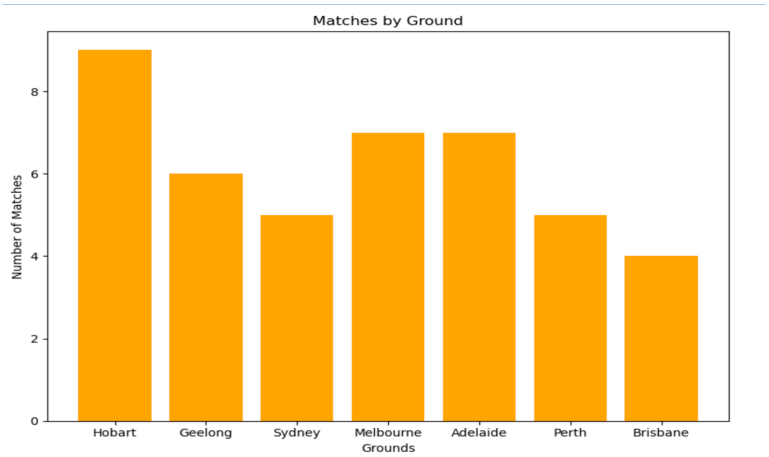

Figure 4.2.8 (b)

Figure 4.2.8(b) shows the venues where Matches were played, and out of all the venues, Hobart has been used for nine matches this year. This suggests that it will be important during the tournament as a main location for people to identify during the event.

4.3.2 Batting Insights:

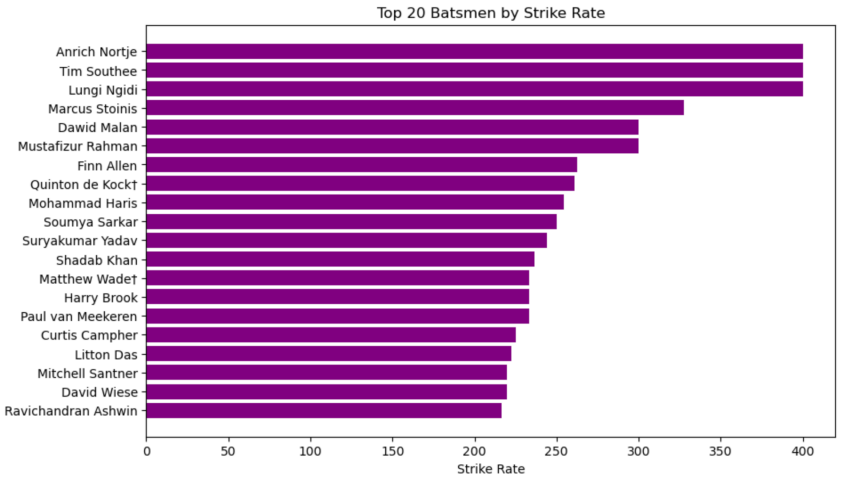

Figure 4.2.9

Figure 4.2.9 represents the batting analysis; two important areas were identified: the strike rates and boundaries that are vital for the player's performance in T20 matches. Thus, Glenn Phillips was characterized by an outstanding strike rate (174.5) and David Miller – by 162.3. The strike rate is one of the factors among the numerous ways in which statistics are employed to tell the story of match performances in cricket. The strike rate for a batsman was calculated using the formula:

Striker rate (SR)= runs made / balls received multiplied by 100

This formula gives the proportion of balls faced that a batsman scores a run and brings out his scoring rate concerning the number of runs scored per 100 balls faced. For example, a strike rate of 174.5, which has been recorded by Glenn Phillip, relates to great scoring power in the T20 system, where quick runs are very important. These figures go a long way in underlining their indispensable capacity to boost scoring, and that will always be a trademark in the shortest versions of the game. The representation of strike rates emphasizes the role of stern batting approaches for receiving those large scores.

4.3.3 Bowling Insights:

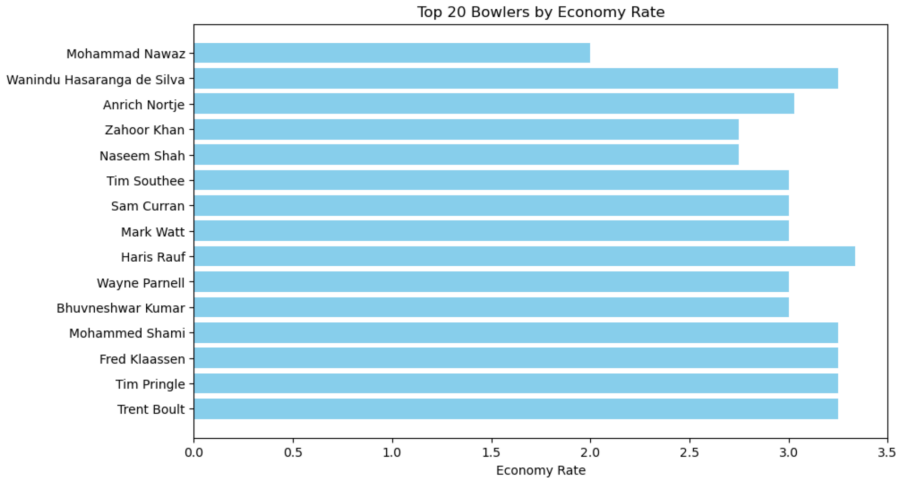

Figure 4.2.10

Figure 4.2.10 refers to the bowling data set some of the key findings were the economy rates of players and their wickets. Wanindu Hasaranga de Silva was the most economical bowler, going with an economy rate of 5.75. The economy rate for a bowler was calculated using the formula:

Economy Rate = Number of runs given by the bowler/number of overs he bowled

This statistic measures the extent of the damage the bowler allows the opposition team to score relative to the number of balls bowled per over in the cricket game. For instance, Wanindu Hasaranga de Silva, who boasts of an economy rate of 5.75, shows his impact in narrowing down, which is important under pressure. This is the reason he has been coming out very handy in terms of denying the opposition side an opportunity to post so many runs on the board. Through the graphical representation of economy rates, the contrast could be effectively made between batsmen-friendly bowlers and bowlers who lose plot under pressure.

4.3.4 Correlation Matrix (Batting Summary):

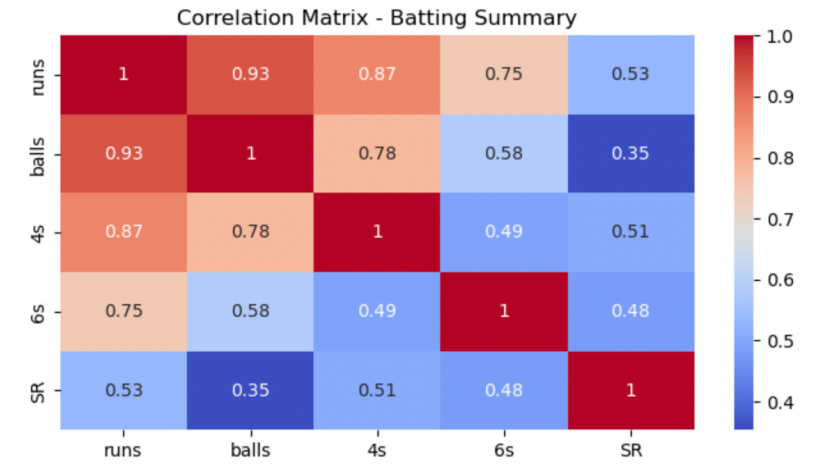

Figure 4.2.11

Strong Positive Correlation: the Runs and Balls faced data has a positive coefficient that shows that there is a trend that the more balls consumed the better the performance in terms of runs made.

4s and 6s Correlation: The frequency count of boundaries (4s and 6s) shows a mere relation with the number of runs, which explains elevated risk-taking.

Strike Rate (SR): SR is determined by a good relationship with the number of 4s and 6s but has a poor relation with balls faced; therefore, SR shows that with a higher strike rate, players score runs quicker.

4.3.5 Boxplot for runs distribution:

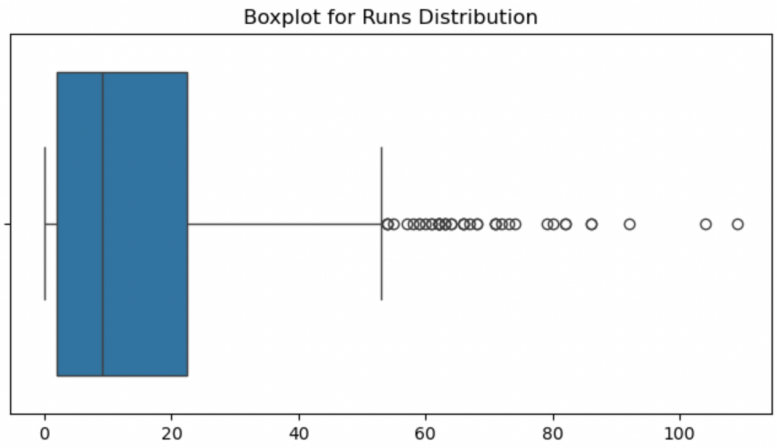

Figure 4.2.12

Outliers Detected: They are skewed a little to the right by a few innings that scored very highly compared to half the median score.

Skewed Distribution: As can be seen, many scores are closer to low values on the graph, which means that the high run-scoring performances are less frequent.

4.3.6 Boxplot for Economy Rate:

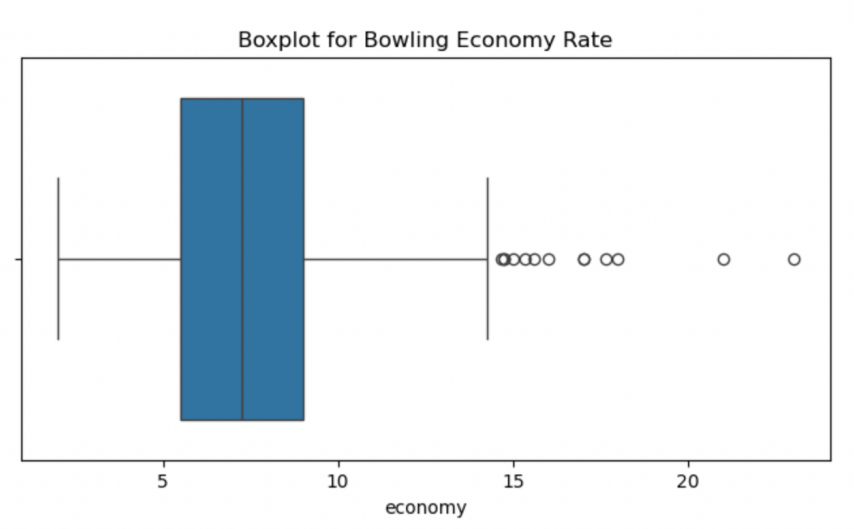

Figure 4.2.13

Skewed Towards Low Economy: Self-organized into clusters, most of the bowlers were found to have a low economy rate, while some bowlers showed a very high economy rate.

4.3.7 Pair plot for Batting Metrics:

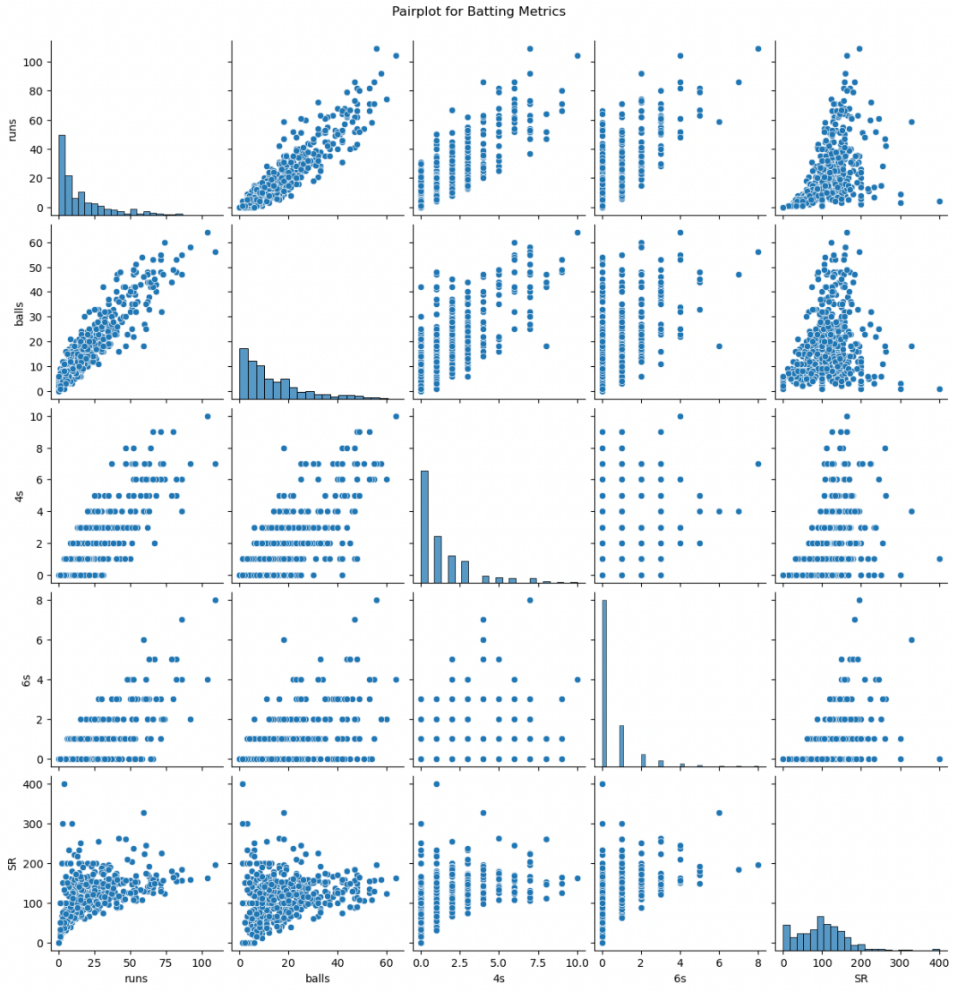

Figure 4.2.14

Diagonal (Histograms):

The diagonal plots on figure 4.3.7 represent the distribution of the individual metrics. For instance, when using a run histogram, we are quickly able to tell that there are many relatively low-scoring innings, but these many high-scoring innings appear anomalous.

Scatter Plots (Off-Diagonal):

Runs vs Balls: There is a positive slope present between the two variables. A general increase in balls faced leads to more runs, therefore IDD shows that longer innings equals a higher overall score.

Runs vs 4s and 6s: The two types of monetary measures were found to have a moderate positive correlation. More number of boundaries tend to receive higher run rates among the players.

Strike Rate (SR) vs Balls: This is evident in a lower correlation than that of ‘balls faced’, suggesting that the strike rate is more inclined by scoring patterns than by mere ‘balls faced’.

Clustering and Patterns: There is grouping shown at the lower end – which is not a surprising finding since most players perform worst at this aspect, with many players scoring few runs and hitting few boundaries in a game. A few ‘outliers’ indicate high scores or risky behavior.

4.4 Power BI DAX:

Power BI was used for further analysis, and DAX queries were applied, the table is given below:

Measures	Description	DAX Formula	Table
Total Runs	Total number of runs scored by the batsman	Total Runs = SUM(fact_batting_summary[runs])	Batting
Total Innings Batted	Total number of innings a batsman got a chance to bat	Total Innings Batted = COUNT(fact_batting_summary[match_id])	Batting
Total Innings Dismissed	To find the number of innings batsman got out	SUM(fact_batting_summary[out])	Batting
Batting Average	Average runs scored in an innings	Batting Avg = DIVIDE([Total Runs],[Total Innings Dismissed],0)	Batting
Total balls Faced	Total number of balls faced by the batsman	total balls faced = SUM(fact_batting_summary[balls])	Batting
Strike Rate	No of runs scored per 100 balls	Strike rate = DIVIDE([Total Runs],[total balls faced],0)*100	Batting
Batting Position	Batting position of a player	Batting Position = ROUNDUP(AVERAGE(fact_batting_summary[batting_pos]),0)	Batting
Boundary %	Percentage of boundaries scored by the Batsman	Boundary % = DIVIDE(SUM(fact_batting_summary[Boundary runs]),[Total Runs],0)	Batting
Avg. balls Faced	Average balls faced by the batter in an innings	AVERAGE(fact_batting_summary[balls])	Batting
Wickets	Total number of wickets taken by a bowler	wickets = SUM(fact_bowling_summary[wickets])	Bowling
balls Bowled	Total number of balls bowled by the bowler	balls Bowled = SUM(fact_bowling_summary[balls])	Bowling
Runs Conceded	Total runs conceded by the bowler	Runs Conceded = SUM(fact_bowling_summary[runs])	Bowling
Bowling Economy	Average number of runs conceded in an over	Economy = DIVIDE([Runs Conceded], ([balls Bowled]/6),0)	Bowling
Bowling Strike Rate	Number of balls bowled per wicket	Bowling Strike Rate = DIVIDE([balls Bowled], [wickets],0)	Bowling
Bowling Average	No. of runs allowed per wicket	Bowling Average = DIVIDE([Runs Conceded],[wickets],0)	Bowling
Total Innings Bowled	Total number of innings bowled by a bowler	Total Innings Bowled = DISTINCTCOUNT(fact_bowling_summary[match_id])	Bowling
Dot Ball %	Percentage of dot balls bowled by a bowler	Dot ball % = DIVIDE(SUM(fact_bowling_summary[zeros]), SUM(fact_bowling_summary[balls]),0)	Bowling
Player Selection	To understand if a player is selected or not	Player Selection = if(ISFILTERED(dim_player[name]),"1","0")	Bowling
Display Text	To display a text of no player is selected	Display Text = if([Player Selection] = "1", " ","Select Player(s) by clicking the player's name to see their individual or combined strength.")	
Color Callout Value	To display a value only when a player is selected	Color Callout Value = if([Player Selection]="0", "#D0CF1D","#1D1D2E")	

4.5 Summary:

The insights gained from this analysis will guide the future development of the project in Power BI. This chapter presented the first results of the research concerning the cricket dataset. Hypotheses generated from EDA were: The team's strategies were understood; This pointed out players' contribution; This presented the dynamics of the entire game. The proposed cleaning, feature engineering, integration, and verification during the data preparation phase involved in this study ensured quality data for analysis.