UNIVERSITI TEKNOLOGI MALAYSIA

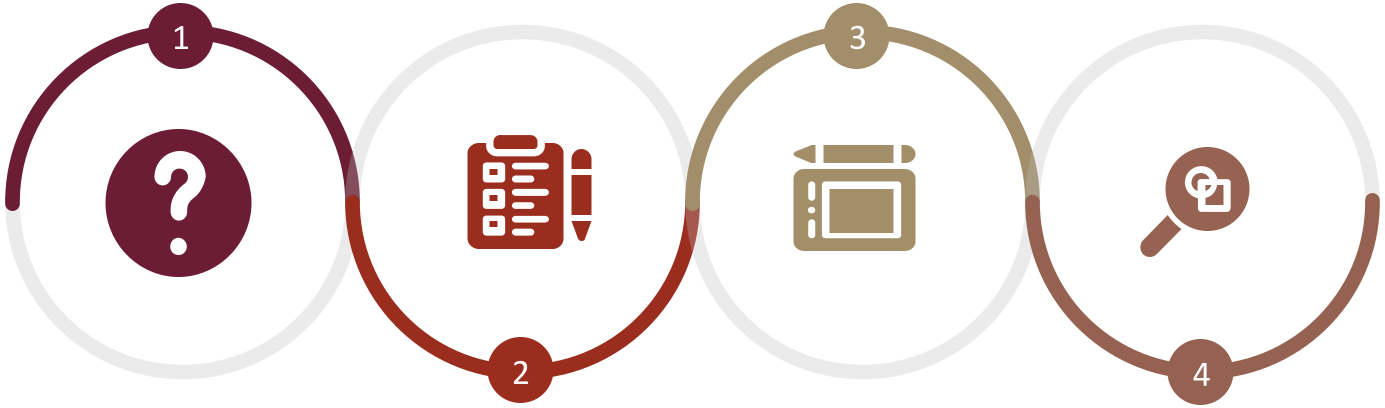
Cricket Data Scraping and Analysis for Robust Data-Driven Decisions

Candidate: Laiba Nadeem
Matric No: MCS241005

Innovating Solutions

Presentation Content

Cricket Data Scraping and Analysis for Robust Data-Driven Decisions



Introduction

Literature Review

Research Methodology

Initial Findings

INTRODUCTION

Cricket Background

Cricket is a bat-and-ball game played between two teams of 11 players. The game is played on a field with a **22-yard pitch** in the center, with wickets at both ends. The **bowler** throws the ball toward the **striker**, who tries to hit it and score **runs** by switching places with the **non-striker**. Runs can also be scored if the ball **reaches the boundary** or if the bowler makes an **illegal delivery**.

Introduction

Problem Statement and Background

PROBLEM BACKGROUND

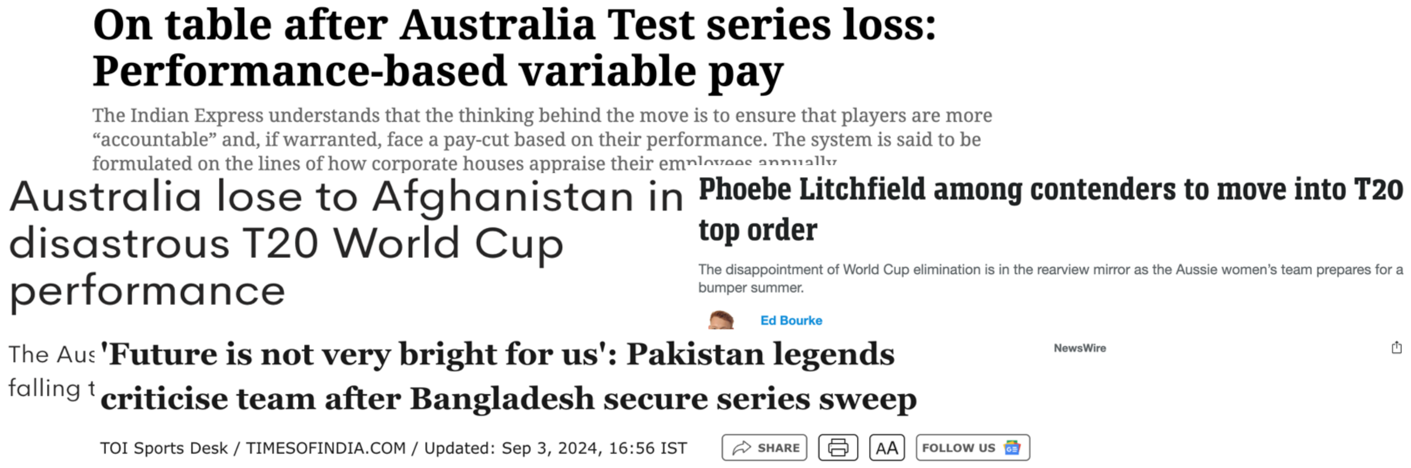
Pakistan faced a disappointing loss in the T20 World Cup, with the head coach citing poor decision-making as a contributing factor. This underscores the impact of selection and strategic choices on match outcomes. Yahoo Sports. (2023).

Rohit Sharma described India's 0-3 loss to New Zealand in a home Test series as a low point in his career. Such outcomes often lead to scrutiny of selection decisions and team strategies. Times of India. (2023).

India's performance in the Border-Gavaskar Trophy was criticized due to inconsistent team selection and imbalanced strategies, leading to a series loss to Australia. Business Standard. (2023).

Reports like the "Sad truth in huge Aussie Test team shake-up" reveal how subjective selection methods often lead to underperforming squads and missed opportunities. Incorporating data analytics could mitigate these issues by objectively analyzing conditions, player fitness, and historical data. Cricket . News.com.au. (2023).

data analytics is transforming cricket, providing teams with critical insights such as player performance trends, opposition weaknesses, and match-specific strategies. It highlights how earlier manual methods were inefficient compared to modern data-driven approaches. Singhal, N., & Jain, A. (2023).



On table after Australia Test series loss: Performance-based variable pay

The Indian Express understands that the thinking behind the move is to ensure that players are more "accountable" and, if warranted, face a pay-cut based on their performance. The system is said to be formulated on the lines of how corporate houses appraise their employees annually

Australia lose to Afghanistan in disastrous T20 World Cup performance

The Aus falling t

Phoebe Litchfield among contenders to move into T20 top order

The disappointment of World Cup elimination is in the rearview mirror as the Aussie women's team prepares for a bumper summer.

Ed Bourke

NewsWire

'Future is not very bright for us': Pakistan legends criticise team after Bangladesh secure series sweep

TOI Sports Desk / TIMESOFINDIA.COM / Updated: Sep 3, 2024, 16:56 IST

PROBLEM STATEMENT

Cricket Data Analysis

Cricket team selection relies on subjective methods, often ignoring available data, limiting decisions based on weather, opponents, and game formats.

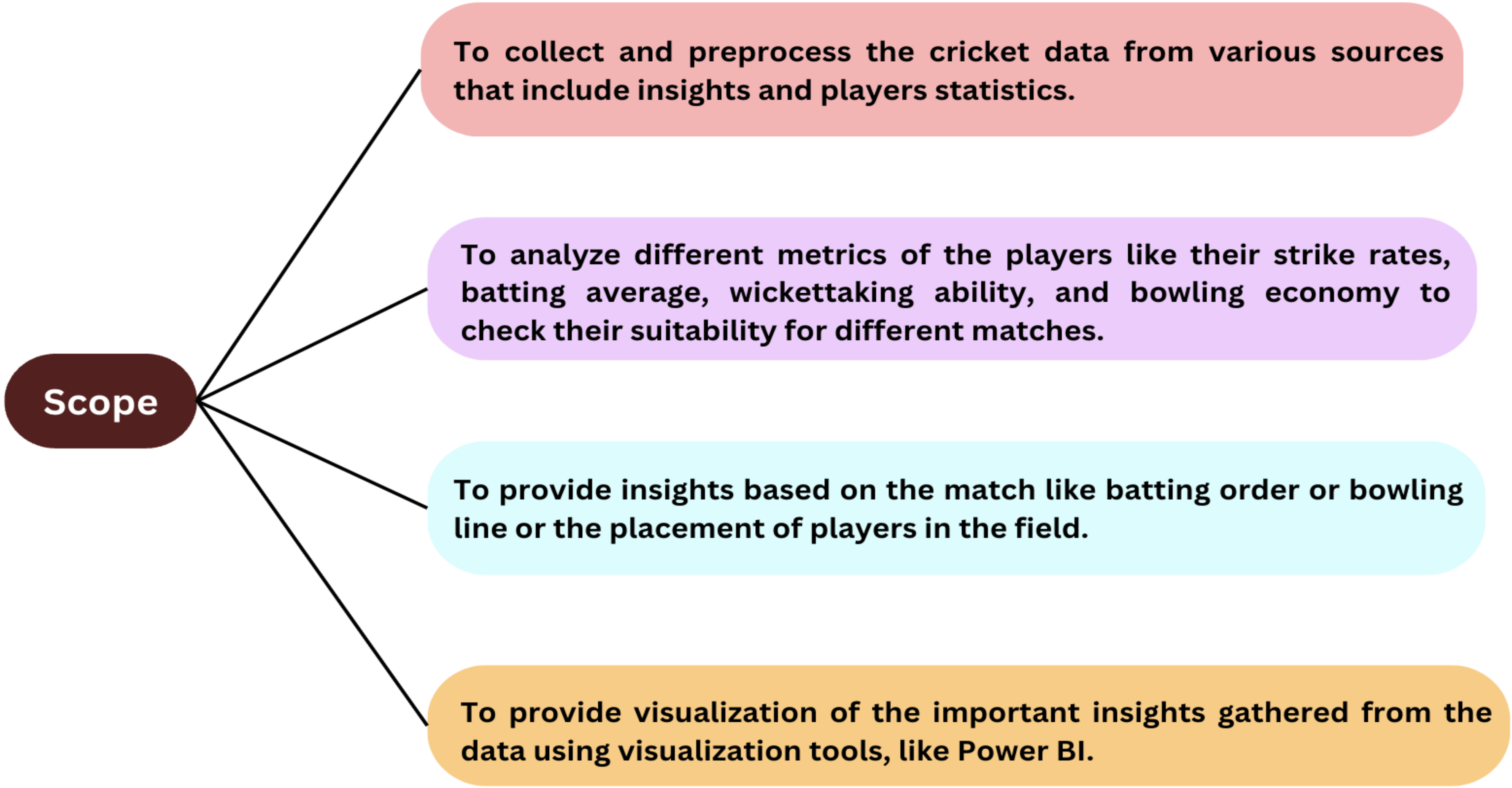
LITERATURE REVIEW

RESEARCH QUESTIONS AND OBJECTIVES

RESEARCH QUESTIONS	RESEARCH OBJECTIVES
How can cricket data be collected, processed, and used to effectively to get data-driven solutions?	To collect and preprocess the cricket data from various sources that include insights and players statistics.
What key metrics are the most important in getting useful insights for team formation?	To analyze different metrics of the players like their strike rates, batting average, wicket-taking ability, and bowling economy to check their suitability for different matches.
How can visualization tools improve the interpretability of the data insights for the team management and coaches?	To provide insights based on the match like batting order or bowling line or the placement of players in the field.
To what extent can Data-driven decisions improve the performance of the team as compared to the traditional method?	To provide visualization of the important insights gathered from the data using visualization tools, like Power BI.

Introduction

RESEARCH SCOPE



Scope

- To collect and preprocess the cricket data from various sources that include insights and players statistics.
- To analyze different metrics of the players like their strike rates, batting average, wickettaking ability, and bowling economy to check their suitability for different matches.
- To provide insights based on the match like batting order or bowling line or the placement of players in the field.
- To provide visualization of the important insights gathered from the data using visualization tools, like Power BI.

RESEARCH GAP

Paper	Key Findings	Gaps/Limitation
Kumar et al., 2019	Machine learning models for player performance prediction in cricket.	**Limited ability to predict player performance under real-time match conditions.**
Sengar et al., 2020	Hierarchical models for player performance prediction in cricket.	**Incomplete or inconsistent real-time data collection during matches.**
Batra et al., 2020	Use of IoT sensors to measure shot quality and player performance.	**Challenges with data normalization and sensor integration.**
Singh et al., 2021	Bayesian models for predicting player performance and team selection.	**Difficulty adjusting for real-time conditions like weather, injuries, or fatigue.**
Verma et al., 2021	Use of predictive modeling based on match data, player rankings, and handedness.	**Incomplete data affecting model accuracy.**
Gupta et al., 2022	Machine learning for cricket team performance benchmarking.	**Need for more sophisticated metrics beyond traditional statistics.**
Patel et al., 2022	Clustering methods for team selection based on complementary abilities.	**Lack of dynamic integration of match data and player injuries.**

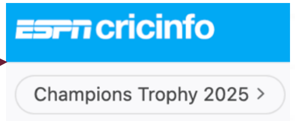
RESEARCH METHODOLOGY

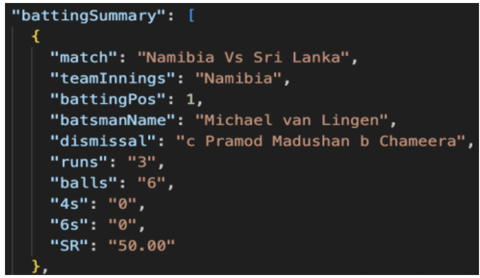
Data Preprocessing

Web Scraper APIs

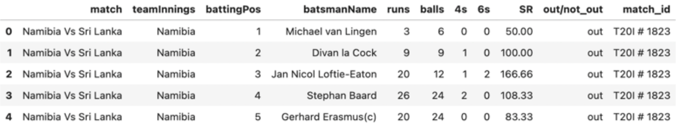
Dedicated endpoints for extracting fresh, structured web data from over 100 popular domains. 100% compliant and ethical scraping.

- ✓ Get structured data in JSON or CSV
- ✓ Bulk request handling, up to 5K URLs
- ✓ Unlimited concurrent scraping tasks
- ✓ Endpoints for crawling and discovery

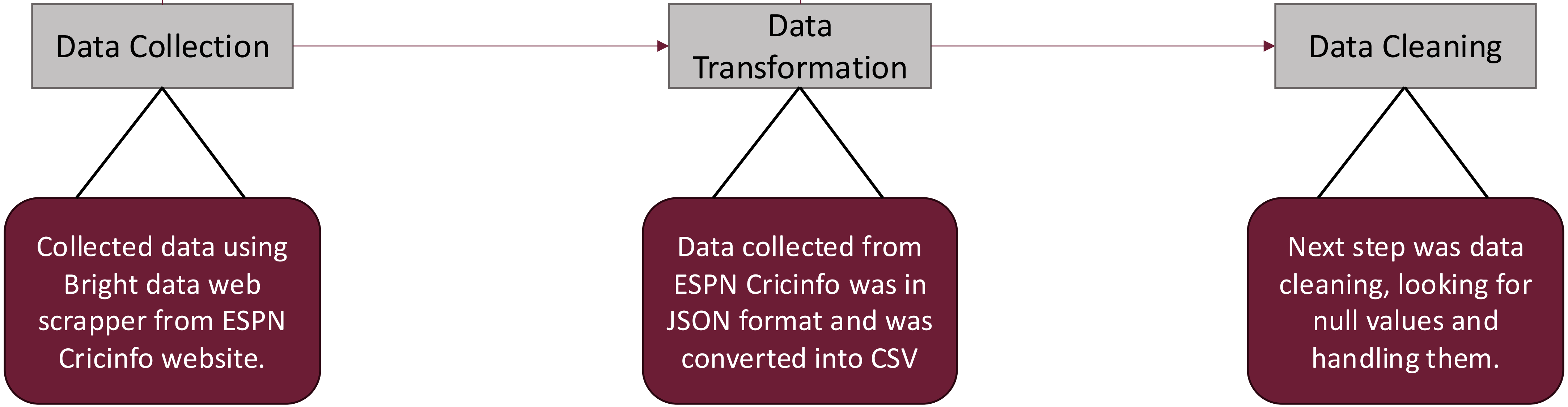




```
"battingSummary": [
  {
    "match": "Namibia Vs Sri Lanka",
    "teamInnings": "Namibia",
    "battingPos": 1,
    "batsmanName": "Michael van Lingen",
    "dismissal": "c Pramod Madushan b Chameera",
    "runs": "3",
    "balls": "6",
    "4s": "0",
    "6s": "0",
    "SR": "50.00"
  },
```



	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



Data Collection

Collected data using Bright data web scrapper from ESPN Cricinfo website.

Data Transformation

Data collected from ESPN Cricinfo was in JSON format and was converted into CSV

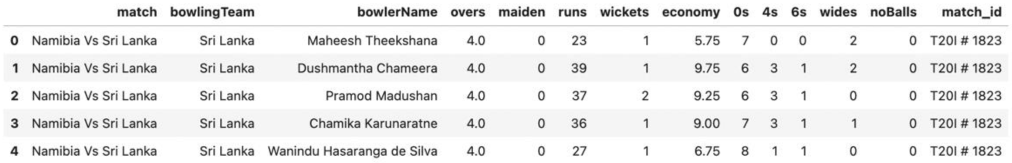
Data Cleaning

Next step was data cleaning, looking for null values and handling them.

INITIAL FINIDINGS

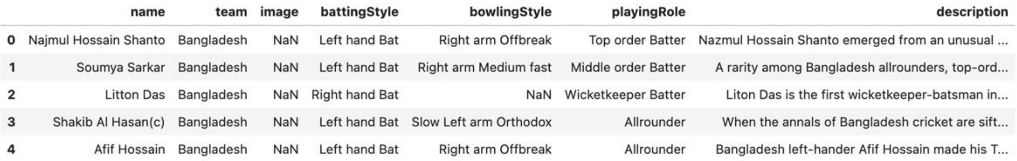
DATASETS

Bowling Data



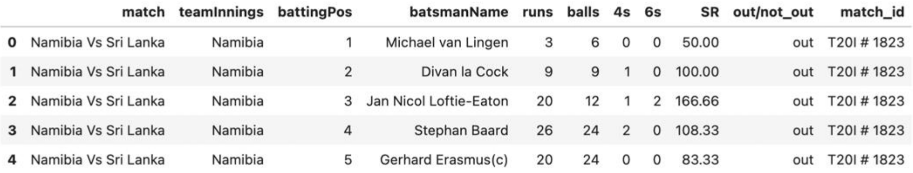
	match	bowlingTeam	bowlerName	overs	maiden	runs	wickets	economy	0s	4s	6s	wides	noBalls	match_id
0	Namibia Vs Sri Lanka	Sri Lanka	Maheesh Theekshana	4.0	0	23	1	5.75	7	0	0	2	0	T20I # 1823
1	Namibia Vs Sri Lanka	Sri Lanka	Dushmantha Chameera	4.0	0	39	1	9.75	6	3	1	2	0	T20I # 1823
2	Namibia Vs Sri Lanka	Sri 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

Players Data



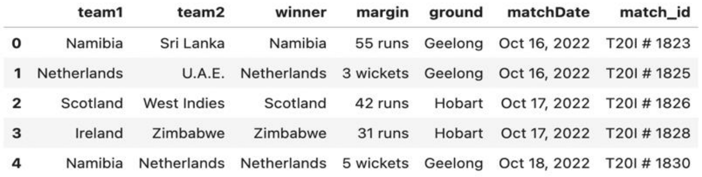
	name	team	image	battingStyle	bowlingStyle	playingRole	description
0	Najmul Hossain Shanto	Bangladesh	NaN	Left hand Bat	Right arm Offbreak	Top order Batter	Nazmul Hossain Shanto emerged from an unusual ...
1	Soumya Sarkar	Bangladesh	NaN	Left hand Bat	Right arm Medium fast	Middle order Batter	A rarity 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...

Batting Data



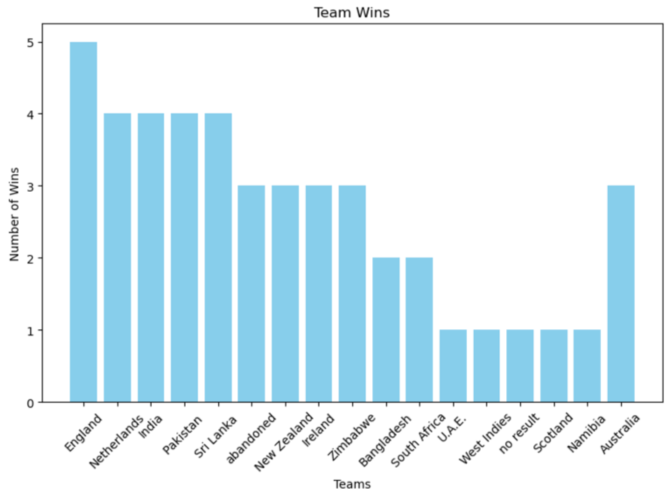
	match	teamInnings	battingPos	batsmanName	runs	balls	4s	6s	SR	out/not_out	match_id
0	Namibia Vs Sri Lanka	Namibia	1	Michael van Lingen	3	6	0	0	50.00	out	T20I # 1823
1	Namibia Vs Sri Lanka	Namibia	2	Divan la Cock	9	9	1	0	100.00	out	T20I # 1823
2	Namibia Vs Sri Lanka	Namibia	3	Jan Nicol Loftie-Eaton	20	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

Match 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

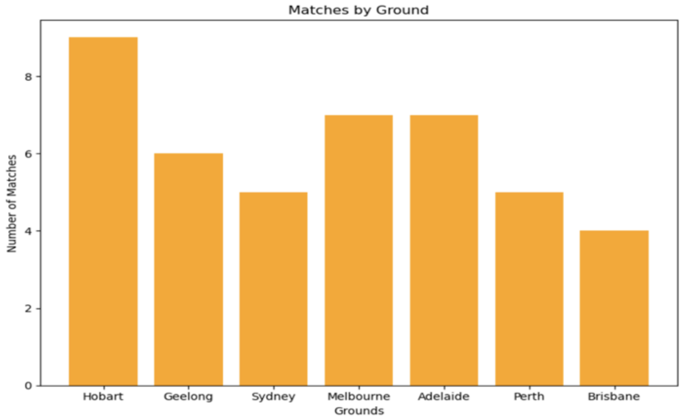
Insights from Data Analysis

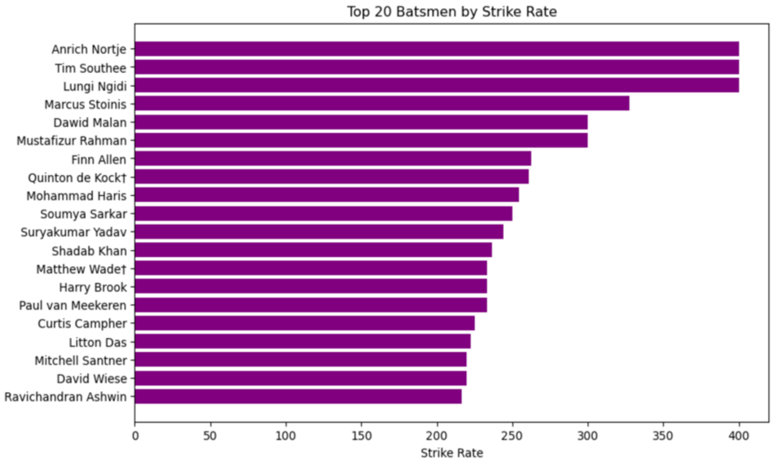


Top team with most wins was England

Most T20's Were played in Hobart

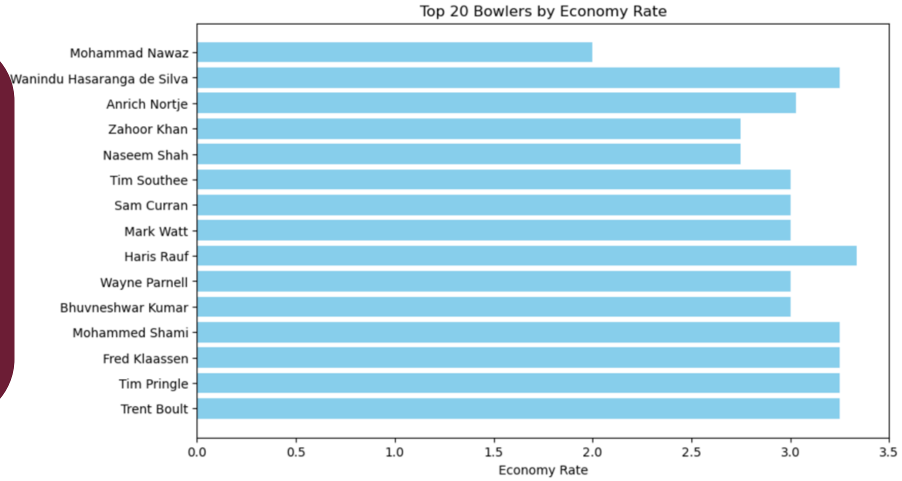
ANALYSES

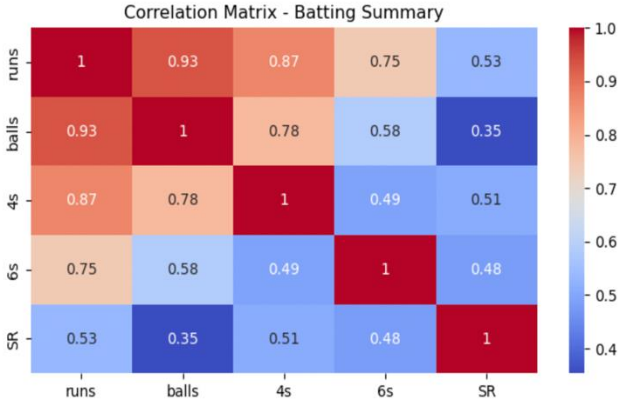




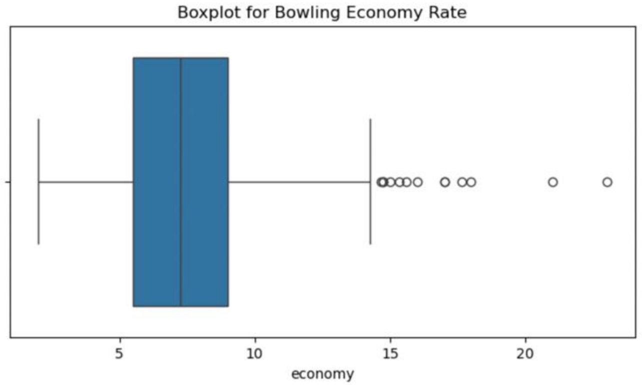
Ordered the batsmen Based on Their Calculated strike rate

Ordered the Bowlers Based on Their Calculated Economy rate

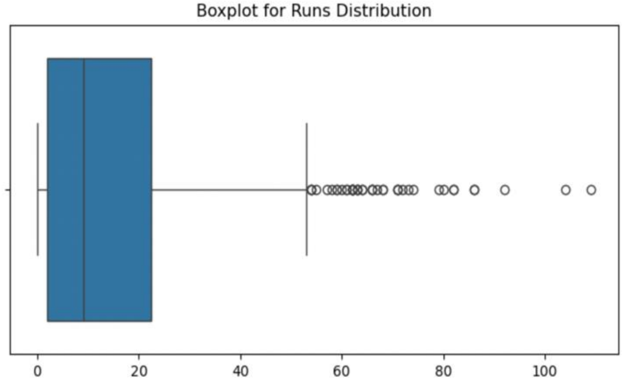




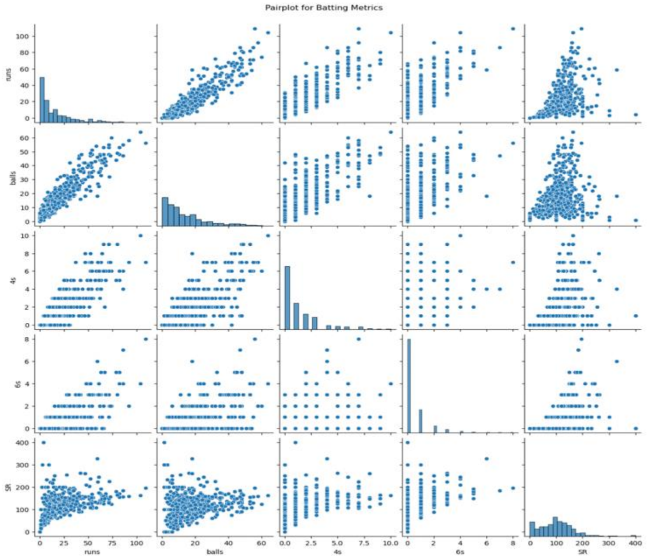
Strong Positive Correlation – More balls faced generally lead to higher runs.**4s and 6s Correlation** – Boundaries slightly impact total runs, indicating risk-taking. Strike **Rate (SR)** – Higher SR means quicker scoring; influenced more by boundaries than balls faced.



Low Economy Clusters – Most bowlers have a low economy rate, but a few have very high rates.

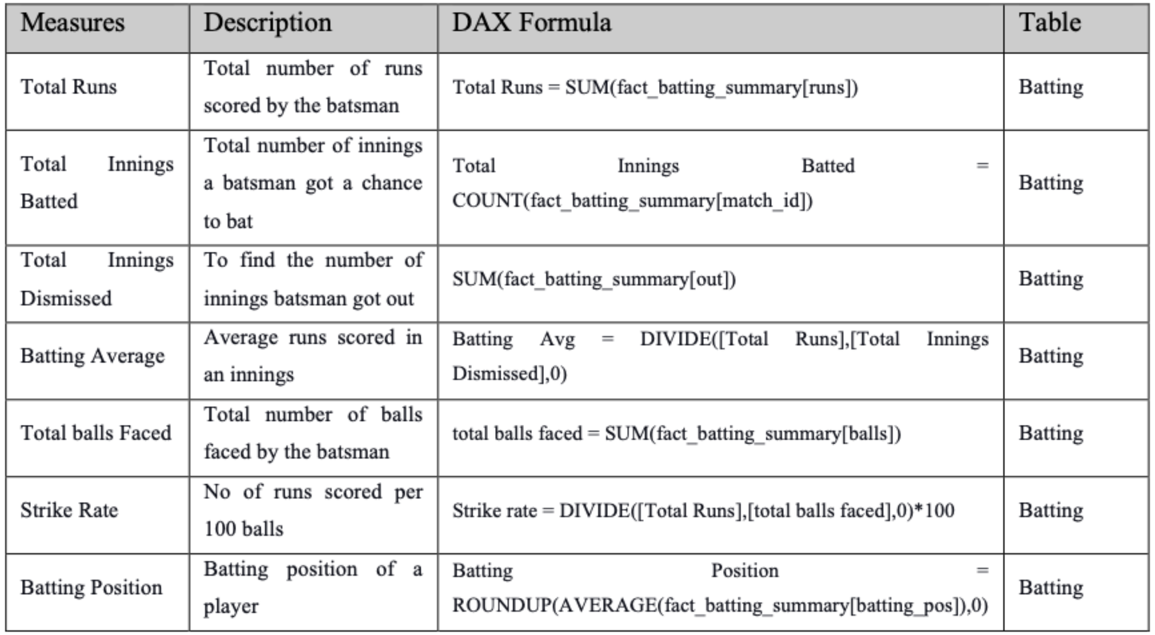


Outliers Detected – A few innings had very high scores, skewing the data right. Skewed **Distribution** – Most scores are low, with fewer high-scoring performances.



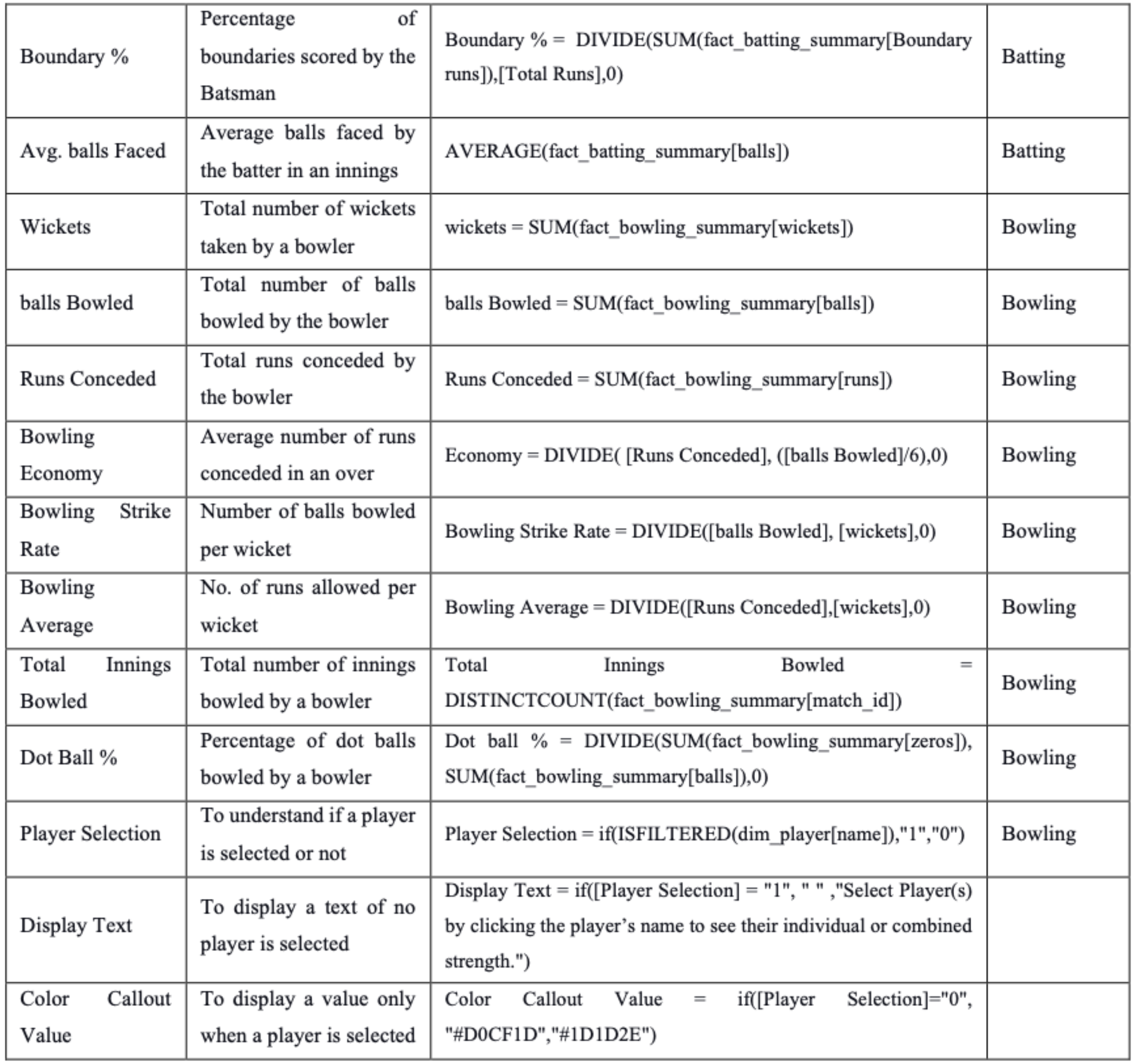
Histograms (Diagonal Plots) – Show individual metric distribution; most innings have low scores, with a few high-scoring outliers. Runs **vs Balls** – More balls faced generally lead to higher runs. Runs **vs 4s & 6s** – More boundaries slightly increase total runs. Strike **Rate vs Balls** – Strike rate depends more on scoring patterns than just balls faced. Clustering – Most players score low; few outliers have high scores, indicating risk-taking.

FUTURE WORK



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

FUTURE WORK:
Using power BI DAX queries and Formulas to develop a dashboard that would give and predict teams on the given demands and requirement of certain match and grounds and against certain team.



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")	

THANK YOU

univteknologimalaysia utm.my utmofficial