Comprehensive Analysis of Cricket Dataset: Insights into Player Performances Across Formats (ODI, T20 and Test)

First A. Jaweria Sohail, Second B. Javeria Hassan, and Third C. Zarnain

Abstract— This report delves into a comprehensive analysis of cricket batting, bowling and fielding data, aiming to uncover valuable insights into player performances across various formats—ODI, T20, and Test. The study employs exploratory data analysis (EDA) techniques to examine key statistical metrics, including runs scored, batting averages, and fielding dismissals. Additionally, probability-based inquiries shed light on the likelihood of players achieving significant milestones such as centuries and notable fielding contributions.

The exploration is conducted on individual and collective player performances, providing a holistic understanding of their strengths and weaknesses in diverse cricketing scenarios. The insights derived from this analysis contribute to a nuanced comprehension of player strategies, form, and adaptability across different formats, ultimately enhancing our appreciation for the dynamic nature of cricket.

Index Terms— Exploratory Data Analysis (EDA), Cricket Statistics, Batting Performance, Fielding Contributions, Probability Analysis, Player Metrics, ODI Format, T20 Format, Test Format, Centuries, Batting Average, Strike Rate, Fielding Dismissals, Data Visualization, Statistical Analysisy

- *F.A. Jaweria Sohail student of Fast University. E-mail:23K8050@nu.edu.pk.*
- *S.B. Javeria Hassan student of Fast University. E-mail:23K8050@nu.edu.pk.*
- *T.C. Jaweria Sohail student of Fast University. E-mail:23K8050@nu.edu.pk.*

1 INTRODUCTION

Cricket, a sport celebrated for its rich history and dynamic gameplay, continues to captivate enthusiasts worldwide. In this report, we embark on a comprehensive analysis of cricket batting and fielding data, seeking to unravel intricate patterns and noteworthy trends within player performances across diverse formats—ODI, T20, and Test.

As cricket evolves with each passing match, the need to understand the nuanced dynamics of player contributions becomes increasingly paramount. This analysis, rooted in Exploratory Data Analysis (EDA) and probability assessments, aims to unearth valuable insights into the factors influencing success on the cricket field.

By scrutinizing individual and collective performances, our exploration extends beyond traditional statistics, offering a holistic perspective on how players navigate the challenges posed by distinct formats. From the pursuit of centuries to the strategic intricacies of fielding, our investigation sheds light on the multifaceted nature of cricket, demonstrating the adaptability and resilience required for success.

Join us on this journey as we delve into the intricate tapestry of cricketing data, deciphering the stories concealed within the numbers and unveiling the essence of player performances in the ever-evolving world of cricket.

2. METHODOLOGY

In our methodology, we executed a robust Exploratory Data Analysis (EDA) on all cricket datasets using powerful data manipulation tools such as pandas and numpy. Our approach involved meticulous handling of missing data and outliers, ensuring the integrity and reliability of our analysis. We adeptly addressed both numeric and non-numeric data, employing appropriate techniques to extract meaningful insights. Statistical analysis techniques were applied to delve into player performances and format-specific dynamics, providing a comprehensive understanding of the dataset. Additionally, our methodology embraced probabilistic assessments, enriching our analysis by evaluating the likelihood of specific events or milestones within the realm of cricket.

3 DATASET

3.1 Dataset Collection

The cricket dataset utilized in this analysis was generously provided by our instructor, forming the foundation for our comprehensive exploration of player performances across diverse formats.

3.2 Dataset Overciew

The datasets for ODI, Test, and T20 formats collectively

consist of an extensive 7,638 rows, capturing a wealth of information across various columns. Key statistics include player details (Player), career span (Span), match statistics (Mat, Inns), bowling metrics (Overs, Mdns, Runs, Wkts, BBI), performance averages (Ave), economy rates (Econ), and strike rates (SR). Additionally, the datasets encompass specific achievements, such as four-wicket (4) and five-wicket (5) hauls. This diverse array of data forms the foundation for a comprehensive analysis, offering insights into player contributions and dynamics across different cricketing scenarios.

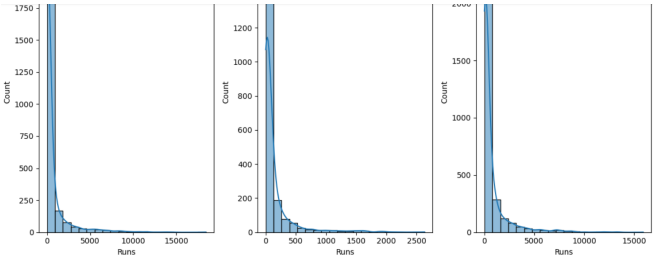
3.3 Data Cleaning

In the initial phase of data cleaning, we meticulously scrutinized the structure of the ODI, T20, and Test datasets individually, focusing on bowling, batting, and fielding categories. Identifying extraneous columns such as 'Unnamed: 0', 'Unnamed: 13', and 'Unnamed: 14', we promptly dropped them to streamline our datasets. Addressing missing values became a priority, and we opted to replace them with zeros, ensuring data completeness. Notably, the 'Span' column, indicating a player's career duration, underwent transformation into an 'Exp' column denoting experience. Further refinement involved parsing 'BBI' columns into separate ones, representing wickets and runs per inning. A critical step involved assessing column data types, revealing some as 'object' types; a conversion to relevant numeric types (integer or float) ensued for enhanced analysis. The comprehensive data cleaning process ensured that no outliers or missing values remained in our datasets, laying a robust foundation for subsequent in-depth analysis.

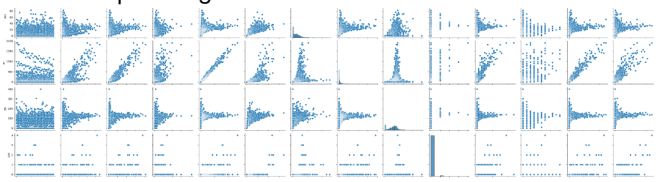
4 Exploartory data Analysis

Following thorough data cleaning, our Exploratory Data Analysis (EDA) commenced with a series of univariate analyses, providing insights into the distribution and characteristics of individual variables. To discern patterns and interrelationships, we plotted pair plots across the datasets, revealing visual representations of associations between various parameters. A pivotal step involved constructing correlation matrices, enabling a comprehensive examination of how variables correlated with one another. In the bowling dataset, a notable finding was the remarkably high correlation coefficient of 0.99 between 'Mat' (Matches) and 'Inn' (Innings), suggesting a strong linear relationship between the two variables. This observation illuminated key insights into player participation and innings played, setting the stage for a nuanced exploration of cricketing dynamics.

Distribution of Batting Runs



Relationship among Numerical Variables



4.1 Univariate Analysis

We conducted univariate analysis for each feature with player data, revealing an initial high trend followed by a gradual decrease (Fig: 1) across the datasets, offering valuable insights into the distribution and patterns of individual player performances.

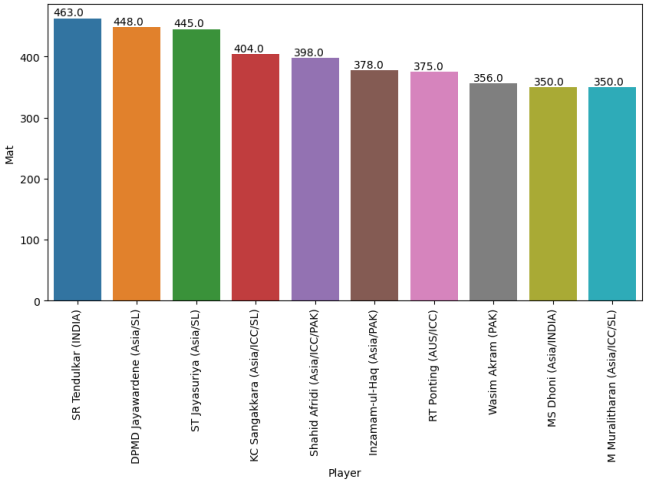


Fig: 1

4.2 Statistical Analysis

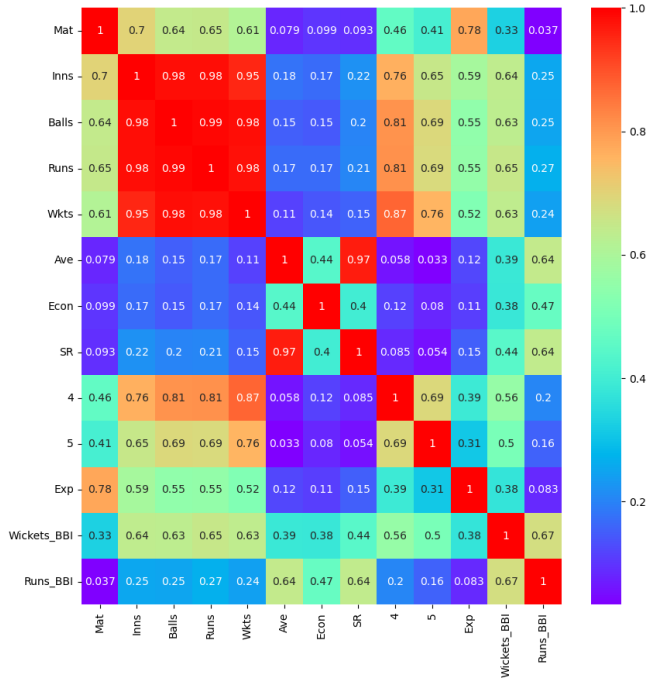
Our statistical analysis delved into each dataset separately, employing measures such as mean, median, and mode (Fig: 2) to gain a comprehensive understanding of the central tendencies within the cricket datasets. By calculating these statistical parameters, we were able to discern the average, middle, and most frequent values for key metrics across batting, bowling, and fielding categories. This methodical approach allowed us to uncover essential insights into the typical values and distributions, contributing to a nuanced exploration of player performances in ODI, T20, and Test formats.

	Mat	Inns	Balls	Runs	Wkts
count	2582	2582	2582	2582	2582
mean	36.03718	19.431448	865.308675	678.481022	20.933772
std	58.2743	39.64081	1866.739974	1407.744194	47.728981
min	1	0	0	0	0
25%	4	0	0	0	0
50%	12	4	130	113	3
75%	41	19	767.25	629.75	18
max	463	372	18811	13632	534

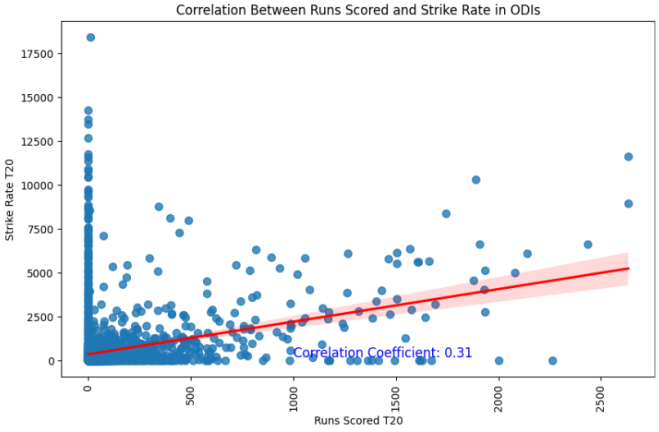
Fig: 2

4.3 Correlation Analysis

Our correlation analysis extended across all datasets, revealing intriguing insights into the relationships between various variables. Notably, the highest correlation within the ODI bowling dataset was observed between 'Runs' and 'Balls,' reaching an impressive coefficient of approximately 0.98 (Fig 3). This strong positive correlation suggests a notable connection between the number of runs scored by bowlers and the total balls bowled during ODI matches. Uncovering such correlations enhances our understanding of the intricate dynamics within the datasets, offering valuable information for further exploration and interpretation of player performances in cricket.



Correlation between Runs Scored and Strike Rates



4.4 Data Distribution Analysis

To unravel the distribution patterns within our datasets, we employed scatter plots as a visual exploration tool. These plots provided a dynamic representation of the relationship between two variables, aiding in the identification of potential trends, clusters, or outliers. The scatter plots allowed us to observe the dispersion and concentration of data points, facilitating a nuanced understanding of the distributional characteristics within cricket batting, bowling, and fielding datasets. This (Fig: 4) visual exploration served as a pivotal step in uncovering hidden patterns and variations, contributing to the overall depth and clarity of our analysis.

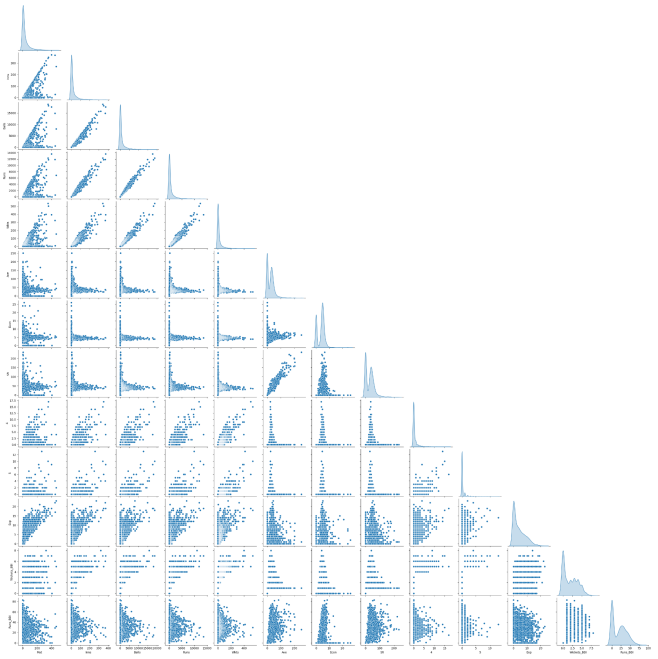
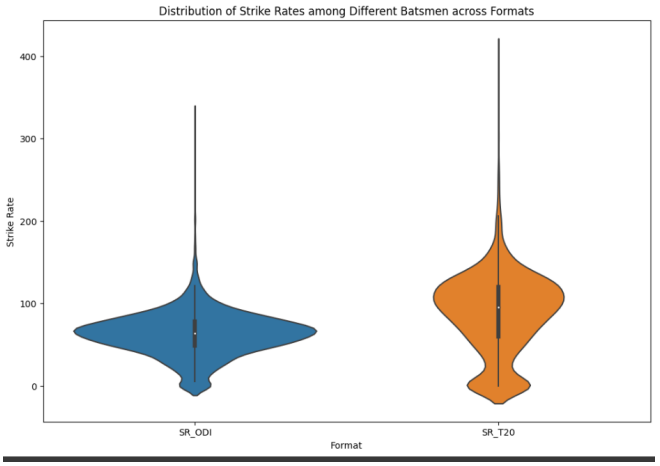
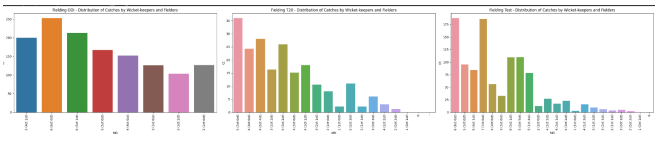


Fig: 4
Distribution of Strike Rates among Different Batsmen across Formats



Distribution of Catches



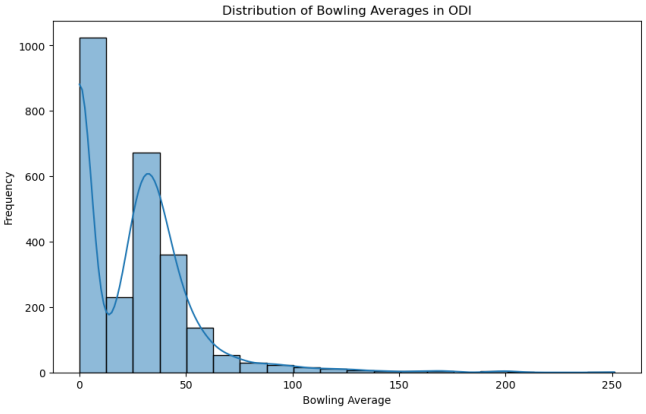
5 Probabilistic Analysis

Our probabilistic analysis delves into various facets of

cricket performances, answering critical questions that illuminate the intricacies of player contributions. We explore the distribution of bowling averages, shedding light on the variability among different players. Moving beyond averages, we investigate the average number of wickets taken by players below specific thresholds, providing nuanced insights into bowling prowess. Economy rates become a focal point, revealing the variation among bowlers and their correlation with wickets taken. Shifting to batting, we explore the distribution of averages among players and decipher the relationship between batting average and the frequency of centuries or half-centuries scored. We further examine the strike rates of different batsmen, probing into their correlation with the total runs scored. This extensive analysis, extending to wicket-keeping and fielding statistics, addresses questions regarding dismissals, catches, and stumpings. Probability calculations offer insights into the likelihood of specific events, such as a player achieving a certain dismissal rate or scoring a century. In summary, our probabilistic analysis enhances our understanding of the diverse dynamics within cricket datasets, providing a foundation for insightful visualizations and interpretations in our report.

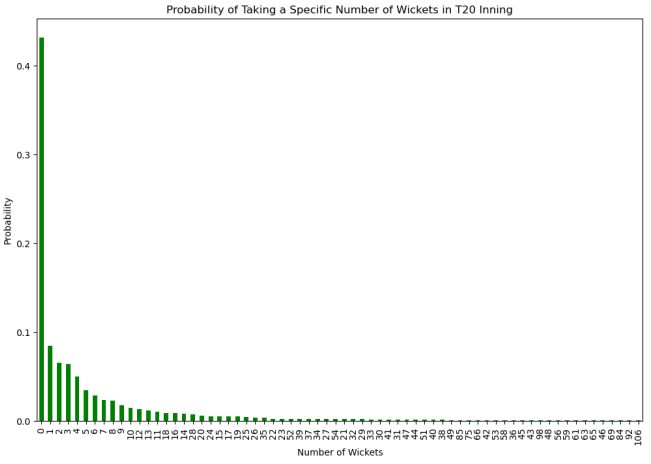
5.1 Bowling Average among Different Players

We plot the histogram to visualize bowling average among different players in ODI matches:



5.2 Probability of Taking Specific Number of Wickets

We plot the histogram to visualize probability of taking specific number of wickets , The highest probability was taking 0 and 1 wickets :

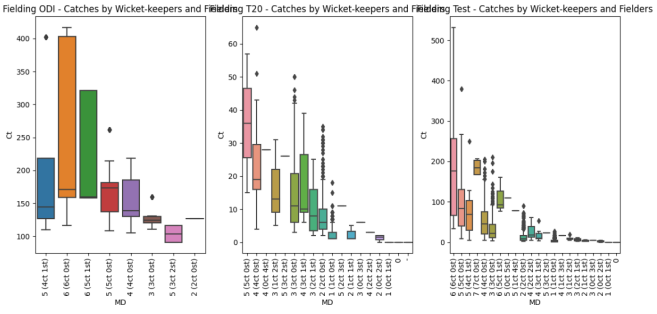


5.3 Best Bowling Performance in ODI

We perform an analysis to check which bowler perform the best by check his number of wickets given by runs :

```
The best bowling performance was by LR Gibbs (WI) in ODI
```

Catches By Wicket Keepers



6 Conclusion

In conclusion, our comprehensive analysis of cricket datasets spanning ODI, T20, and Test formats has unearthed valuable insights into the multifaceted realm of player performances. Through meticulous data cleaning, exploratory data analysis, and probabilistic assessments, we navigated the intricate landscape of batting, bowling, and fielding statistics. From uncovering distribution patterns and correlations to addressing nuanced questions about averages, wickets, and centuries, our exploration has provided a holistic understanding of cricket dynamics. The identification of key trends, such as the strong correlation between runs and balls in ODI bowling, serves as a testament to the depth of our analysis. These findings not only contribute to the broader field of cricket analytics but also offer a nuanced appreciation for the strategic nuances and adaptability required for success in this dynamic sport.

Acknowledgment

I extend my heartfelt gratitude to Javeria Hassan, Jaweria Sohail, and Zarnain Maryam for their unwavering

support and valuable contributions throughout this analytical journey. Their dedication and collaboration significantly enriched the depth and quality of this project. Special appreciation is owed to Dr. Fahad Samad for his exemplary supervision and guidance, which played a pivotal role in shaping the direction of this endeavor. Their expertise and encouragement have been invaluable, and I am sincerely thankful for the mentorship and insights they generously shared.

First A. Jaweria Sohail (23K-8050) Jaweria Sohail, a dynamic software engineer, is an alumna of Iqra University, where she honed her skills and acquired a solid foundation in the field. Currently, she serves as a dedicated software engineer at Softech Worldwide, showcasing her proficiency and commitment to the ever-evolving landscape of technology. Jaweria's journey from academia to the professional realm is a testament to her relentless pursuit of excellence and her ability to leverage her education for real-world impact. Her role at Softech Worldwide reflects her expertise in addressing complex challenges and contributing to innovative solutions within the software industry. Jaweria's passion for technology and her professional achievements underscore her commitment to pushing the boundaries of what's possible in the field of software engineering.

Second B. Javeria Hassan (23K-8019) Javeria Hassan, a distinguished software engineer, completed her undergraduate studies at PAF-KIET, showcasing exceptional acumen and dedication to her field. Currently serving as an instructor in a non-profit organization, Javeria generously imparts her profound knowledge through dynamic software courses, contributing significantly to the educational landscape. Beyond her instructional role, she also excels as a Python developer at AiSol Robotics Solutions, where her technical prowess and innovative mindset contribute to cutting-edge advancements in the field of robotics. Javeria's unwavering commitment to both education and technological innovation exemplifies her passion for making meaningful contributions to the realm of software engineering and artificial intelligence.

Third C. Zarnain Maryam Awan (23K-7702) Zarnain Maryam Arwan, a proficient software engineer, proudly holds her degree from Virtual University, reflecting her dedication to academic excellence. With a solid educational foundation, Zarnain has embarked on a promising career in the dynamic field of software engineering. Her journey is marked by a commitment to continuous learning and a passion for staying abreast of cutting-edge technologies. As a software engineer, Zarnain channels her skills to navigate the intricate challenges of the industry, bringing innovation and efficiency to her work. Her educational background from Virtual University serves as a testament to her adaptability and resilience in the face of evolving technological landscapes. Zarnain Maryam Arwan's contributions to the software engineering field exemplify her drive to make a meaningful impact through her technical expertise and problem-solving abilities..