

A Project Report

on

Kohli's Record-breaking Journey: A Statistical Perspective

BACHELOR OF SCIENCE

in Applied Statistics and Data Science

by

SNEHA KUMARI 039

PRACHI 032

ADITYA MAKKAR 028

DIVYA KANODIA 062

PREKSHA UPPAL 040

Under the supervision of

Dr. Vidyा Yerneni



Symbiosis Statistical Institute

October 2024

ACKNOWLEDGEMENT

An undertaking is never a result of a solitary individual; rather it bears the engravings of various individuals who specifically or by implication helped in finishing that venture. We would bomb in my obligations on the off chance that we don't let out the slightest peep of gratitude to every one of the individuals who helped us in finishing this task of our own.

Before we start with the details of my projects, we would like to add a few heartfelt words for the people who were part of my project in numerous ways, the people who gave me their immense support right from the initial stage.

As a matter of first importance, we are amazingly appreciative of **Dr. Vidya Yerneni** for her direction, consolation, smooth feedback, and tutelage throughout this task notwithstanding her to a great degree occupied timetable.

We also heartily thank our friends who greatly helped us in our project work. Without them we would never have gained the actual problem set solutions that we faced.

Last but not the least, we heartily thank our respected director **Dr. Sharvari Shukla** who kept on pushing our limits and taught us to be positive in every way.

Prachi 032

Sneha Kumari 039

Aditya Makkar 028

Preksha Uppal 040

Divya Kanodia 062

Abstract

This project presents a data-driven analysis of a cricket player's career performance, spanning multiple years, using detailed statistical metrics to examine trends, consistency, and match impact. By focusing on year-by-year player statistics, this report provides insights into batting effectiveness, scoring patterns, and the player's impact on match outcomes. Key performance indicators include runs scored, highest scores, averages, strike rates, and balls faced, along with an assessment of key milestones such as centuries, fifties, and "Player of the Match" awards.

Descriptive statistics, such as mean, median, variance, skewness, and kurtosis, offer a comprehensive understanding of the player's scoring patterns and performance volatility. Through statistical correlation analysis, relationships between key metrics (e.g., runs scored, balls faced, and match frequency) are explored, highlighting trends in the player's scoring ability and consistency. Notable findings reveal a strong positive correlation (0.976) between runs scored and balls faced, reflecting a linear relationship, while the weaker correlation between match frequency and run total indicates consistent scoring independent of match frequency.

This analysis provides a deep dive into the player's peak performance periods and highlights performance shifts over time, with standout years such as 2016 marked by high scoring, elevated strike rates, and increased match-winning contributions. Overall, the project offers valuable insights into the player's career trajectory and adaptability, using both quantitative metrics and qualitative observations to provide a holistic evaluation of their cricketing prowess. The findings underscore the player's contributions and adaptability, presenting a comprehensive evaluation that contributes to understanding their career trajectory and overall cricketing impact.

TABLE OF CONTENTS

Acknowledgement.....	2
Abstract.....	3
Table of contents.....	4
1. Data Description.....	5
2. Dataset.....	6
3. Data Cleaning and Preparation.....	7
4. Statistical Methods.....	8
4.1 Descriptive Statistics	
4.2 Inferential Statistics	
5. Visual Presentation of Data.....	12
5.1 Runs over the years	
5.2 Strike rate trend	
5.3 Centuries(100s) and fifties(50s)	
5.4 Fours(4s) and sixes(6s)	
5.5 Matches vs. not outs	
5.6 Correlation between balls faced and runs	
5.7 Correlation between matches and runs	
5.8 Correlation between matches and not outs	
6. Trend Analysis.....	17
7. Conclusion.....	17
8. Limitations.....	17
9. References.....	18

Data description

Virat Kohli has created history with his achievements and success in his cricket career. He was the youngest team captain and also led his team to 40 victories, only 17 loses and 11 ending in a draw. His home and away record made him one of India's most successful captains in the red-ball cricket.

Kohli made his international debut for India in 2008 and quickly rose to prominence due to his aggressive batting style, exceptional consistency, and remarkable fitness.

The dataset spans 17 years and includes several key cricket performance metrics. Here are the data fields used:

Year: Indicates the year of performance.

Matches: Total matches played in each year.

Not Out (N.O): The number of times the player was not out at the end of an innings.

Runs: Total runs scored across all matches played in a year.

High Score (HS): The highest score in an innings in that particular year.

Average (Avg): Calculated as the total runs divided by the number of dismissals.

Balls Faced (BF): Total number of balls faced.

Strike Rate (SR): Runs per 100 balls faced.

100s & 50s: Number of centuries (100+) and half-centuries (50+) scored in a year.

Fours (4s) & Sixes (6s): Total boundaries hit by the player.

Player of the Match (POTM): The number of Player of the Match awards.

DATASET

The dataset contains performance metrics for 17 years, from 2008 to 2024. Each row represents the player's performance in a given year. Metrics such as the number of matches, total runs, and strike rate are essential for determining the player's effectiveness and consistency over the years.

Year	Matches	N.O.	Runs	High score	Avg	BF(balls faced)	Strike rate	100s	50s	4s	6s	Player of the match
2024	15	3	741	113	61.75	479	154.7	1	5	62	38	2
2023	14	2	639	101	53.25	457	139.82	2	6	65	16	2
2022	16	1	341	73	22.73	294	115.99	0	2	32	8	1
2021	15	1	405	72	28.92	339	119.46	0	3	43	9	0
2020	15	4	466	90	42.36	384	121.35	0	3	23	11	1
2019	14	0	464	100	33.14	328	141.46	1	2	46	13	1
2018	14	3	530	92	48.18	381	139.1	0	4	52	18	0
2017	10	0	308	64	30.8	252	122.22	0	4	23	11	0
2016	16	4	973	113	81.08	640	152.03	4	7	83	38	5
2015	16	5	505	82	45.9	386	130.82	0	3	35	23	1
2014	14	1	359	73	27.61	294	122.1	0	2	23	16	0
2013	16	2	634	99	45.28	457	138.73	0	6	64	22	3
2012	16	2	364	73	28	326	111.65	0	2	33	9	0
2011	16	4	557	71	46.41	460	121.08	0	4	55	16	2
2010	16	2	307	58	27.9	212	144.81	0	1	26	12	0
2009	16	2	246	50	22.36	219	112.32	0	1	22	8	0
2008	13	1	165	38	15	157	105.09	0	0	18	4	0

DATA CLEANING AND PREPARATION

Data cleaning and preparation are essential steps in ensuring the accuracy and reliability of the analysis. For this dataset, which includes various cricket performance metrics from 2008 to 2024, the following cleaning and preparation steps were carried out:

1. Handling Missing Data- An initial check was performed to detect missing values using standard techniques. Fortunately, the dataset did not contain any missing values across columns like Matches, Runs, Balls Faced, Strike Rate, and others, ensuring completeness for analysis.

2. Dealing with Outliers- Key outliers were observed in years like 2016 (973 runs) and 2024 (741 runs), where the player performed significantly better than in other years. Additionally, 2012 was flagged as a year of lower performance. These outliers were not removed, as they represent important career milestones and help illustrate key trends in the player's performance.

3. Ensuring Data Consistency- Data consistency was checked by verifying relationships between fields such as Matches, Runs, Balls Faced, and Strike Rate. No inconsistencies were found.

4. Data Type Verification- Each column was checked to confirm that the correct data types were used. Numeric fields like Runs, Balls Faced, and Strike Rate were stored as integers or floats, while Year was treated as an integer. No data type conversions were required.

- **Final Data Preparation-** After completing the cleaning steps, the dataset was confirmed ready for analysis. No missing data, inconsistencies, or duplicates were found, and outliers were flagged but retained. With the data cleaned, the next step is to dive into the exploratory analysis to uncover trends and insights.

This concise cleaning process ensures that the dataset is reliable and well-prepared for EDA.

Statistical Methods

DESCRIPTIVE STATISTICS- Descriptive statistics refers to the branch of statistics that summarizes and describes the features of a dataset. It provides simple quantitative summaries about the sample and the measures. Descriptive statistics help in understanding the basic characteristics of data without making inferences or predictions.

❖ Measure Of Central Tendency

Measures of central tendency are statistical metrics that describe typical value of a dataset. They provide a single representative value that summarizes a collection of numbers. The main measures include the mean (average), median (middle value), and mode (most frequent value), each serving to highlight different aspects of the data distribution.

1. **MEAN-** The mean, or arithmetic average, is calculated by summing all the values in a dataset and dividing by the total number of values. It is applicable for quantitative data only. The mean is commonly used to summarize data and is sensitive to extreme values (outliers), which can affect its representation of the dataset.
2. **MEDIAN-** The median is the middle value of a dataset when the values are arranged in ascending or descending order. If there is an even number of observations, it is the average of the two middle values. The median is a measure of central tendency that is less affected by outliers, providing a better indication of the centre in skewed distributions. It is also known as positional average.
3. **MODE-** The mode is the value that appears most frequently in a dataset. A dataset may have one mode (unimodal), two modes (bimodal), multiple modes (multimodal), or no mode at all. The mode is useful for identifying the most common value in categorical data and for highlighting trends in quantitative data, though it may not always represent the overall distribution effectively.

Year	Match	N.O.	Runs	HS	Avg	BF	SR	100s	50s	4s	6s	POM
MEAN	14.823	2.1764	470.82	73.9	38.862	356.764	128.98	0.4705	3.2352	41.4	16	1.058
MEDIAN	15	2	464	72	33.14	339	122.22	0	3	35	13	1
MODE	16	2	#N/A	73	NA	457	NA	0	2	23	16	0

❖ Measures of Dispersion

Measures of dispersion, also known as measures of variability or spread, quantify the extent to which data points in a dataset differ from each other and from the central tendency (mean, median, or mode). They provide insight into the distribution and spread of data, helping to assess consistency and reliability.

1. **QUARTILES-** They are specific types of percentiles that divide a dataset into four equal parts, providing insights into the spread and distribution of the data.
 - First Quartile (Q1): The median of the lower half of the dataset (25th percentile).
 - Second Quartile (Q2): The median of the dataset is also known as second quartile (50th percentile).
 - Third Quartile (Q3): The median of the upper half of the dataset (75th percentile).
2. **VARIANCE** – It is a statistical measure that quantifies the degree of spread or dispersion in a dataset. It assesses how far each data point in the set deviates from the mean, providing insight into the overall variability within the data. For each data point, you subtract the mean, square the result (to eliminate negative values), and then average those squared differences by dividing by the total number of observations. 2nd Quartile is also known as variance.
3. **STANDARD DEVIATION (SD)**- It is a fundamental statistical measure that quantifies the amount of variation or dispersion in a dataset. It indicates how much individual data points deviate from the mean (average) of the dataset.

The standard deviation is the square root of the variance, which measures the average squared deviation of each data point from the mean. It provides a measure of spread in the same units as the original data, making it interpretable in the context of the data.

- A **low standard deviation** indicates that the data points are clustered closely around the mean, suggesting low variability.
- A **high standard deviation** indicates that the data points are spread out over a wider range of values, indicating greater variability.

	Match	N.O	Run	HS	Avg	BF	SR	100s	50s	4s	6s	POM
Q1	14	1	341	59.5	27.9	294	119.4	0	2	23	9	0
Q3	16	3	557	92.5	46.41	457	139.8	0	4	55	18	2
Q2	2.52	2.1 5	396 73. 65	1721 .765	276.4 551	14305. 316	221.1 984	1.139 706	3.816 176	370.0 147	93.8 75	1.933 824
SD	1.59	1.4 67	199 .1	41.4 9	16.62	119.6	14.87	1.067	1.953	19.23	9.68	1.390

❖ SKEWNESS

Skewness is a statistical measure that quantifies the asymmetry of a probability distribution around its mean.

Types of Skewness:

1. Positive Skewness (Right Skew)

- In a positively skewed distribution, the tail on the right side (higher values) is longer or fatter than the left side. Values greater than 0 indicate a right-skewed distribution.
- Most data points are concentrated on the left, with a few larger values on the right. Values less than 0 indicate a left-skewed distribution.

2. Negative Skewness (Left Skew)

- In a negatively skewed distribution, the tail on the left side (lower values) is longer or fatter than the right side.
- Most data points are concentrated on the right, with a few smaller values on the left.

❖ KURTOSIS- Kurtosis measures the "tailedness" of a data distribution,

indicating how much data is in the tails versus the centre.

- A distribution with a kurtosis value around 3 is considered mesokurtic.
- A distribution with kurtosis greater than 3 is termed leptokurtic.
- A distribution with kurtosis less than 3 is called platykurtic. It has lighter tails and a flatter peak, indicating a lower likelihood of extreme values and a more uniform spread of data.

	Match	N.O	Runs	HS	Avg	BF	SR	100s	50s	4s	6s	POM
SKEWNESS	1.892	0.328	0.919	0.2967	0.982	0.5131	0.226	2.705	0.3675	0.635	1.408	1.62
KURTOSIS	4.45	0.71	1.19	0.7093	1.15	0.5795	1.081	7.661	0.5475	0.576	1.587	2.866

INFERENTIAL STATISTICS - It is a branch of statistics that focuses on making conclusions about a larger group based on a smaller subset of data. It allows us to analyze and interpret data in a way that goes beyond mere description, helping us to make predictions and test hypotheses.

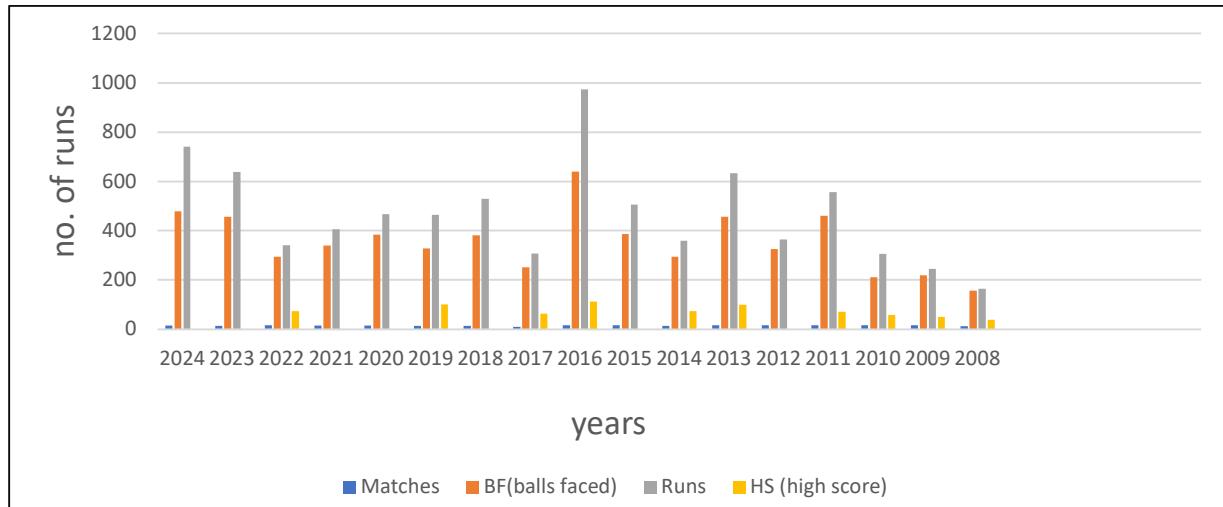
- **CORRELATION** - Correlation is a statistical concept that describes the relationship between two variables. In simple terms, it tells us whether and how strongly two things are related to each other. If the two-variable deviate in the same direction, i.e., if the increase or decrease in one, results in a corresponding increase or decrease in the other, correlation is said to be direct or positive. But if they constantly deviate in the opposite direction, i.e., if increase or decrease in one causes a decrease or increase in the other, correlation is said to be diverse or negative. Correlation is said to be perfect if the deviation in one variable is followed by a corresponding and proportional deviation in the other.

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

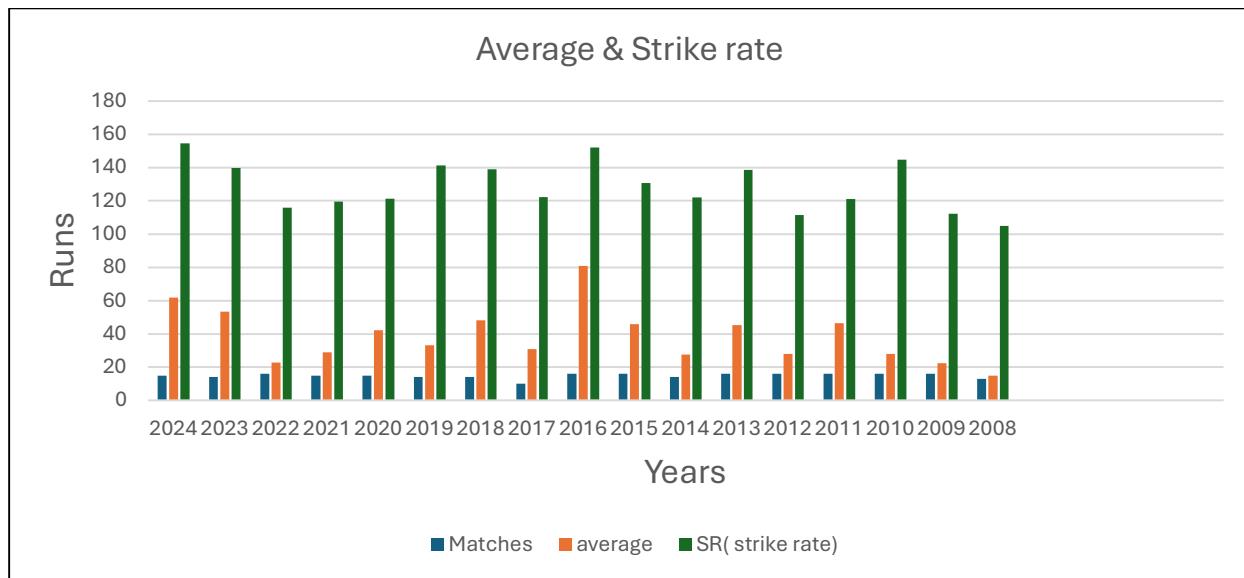
Correlation	Value	Interpretation
correlation b/w bf and runs	0.976804	positive and strong relation
correlation b/w matches and runs	0.2915	positive but weak relation
Correlation b/w matches and not out	0.549645	moderate positive relation

VISUAL PRESENTATION OF DATA

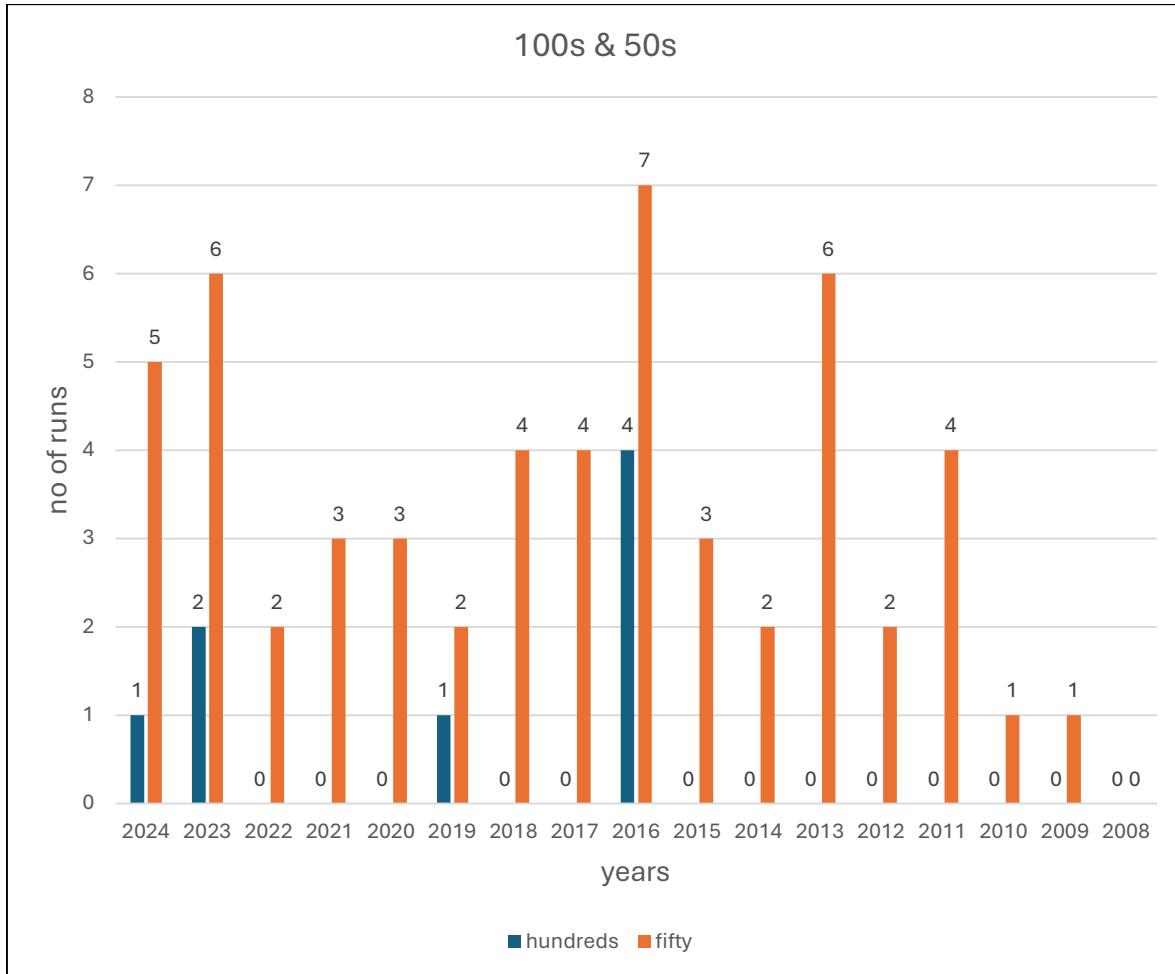
4.1. Runs Over the Years – The player scored their highest runs in 2016 with 973 runs across 16 matches (60.81 runs per match). In 2024, they scored 741 runs in 15 matches, showing a strong recent performance. Low-performing years, like 2008 with 165 runs in 13 matches, reflect a struggling form. This graph shows a peak performance period between 2016 and 2024, with consistently high runs. The highest score recorded was 113 in 2024, contributing to a total of 741 runs.



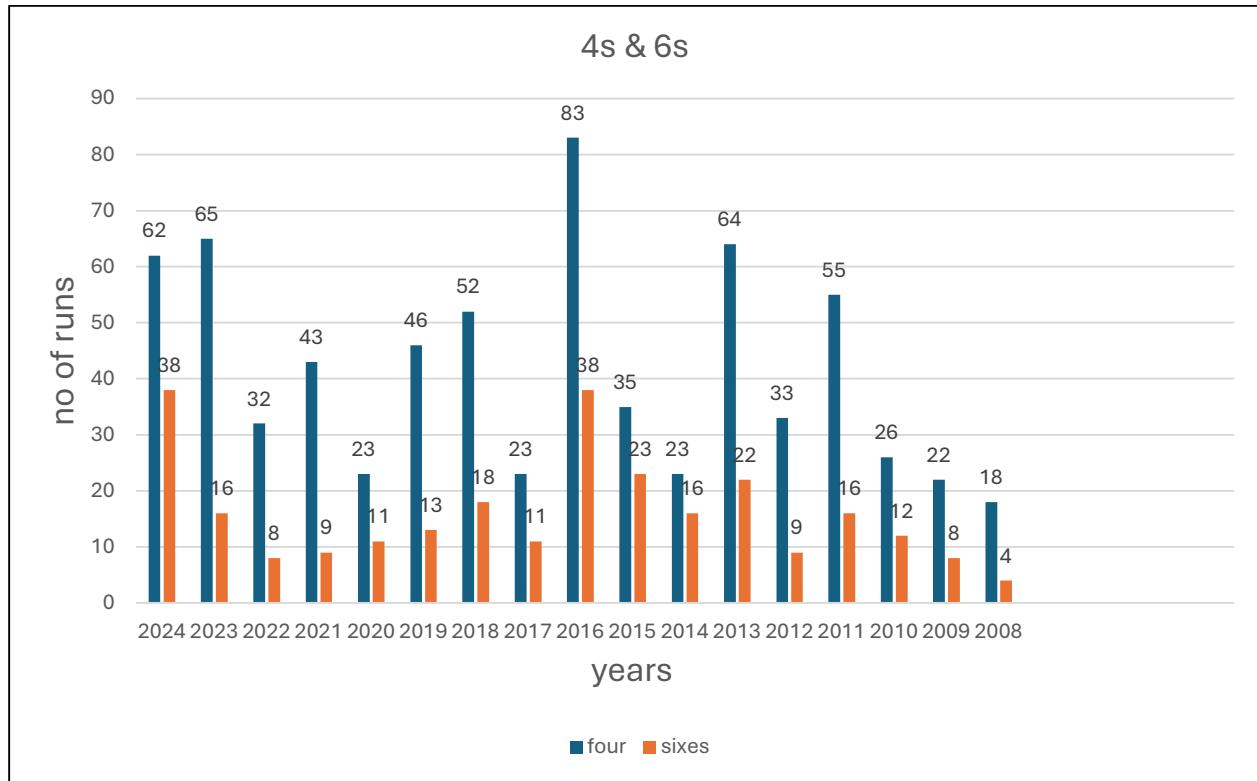
4.2. Strike Rate Trend- In 2016, the player achieved a peak batting average of 81.08 with a strike rate of 152.03. The 2024 season showed an average of 61.75 and a strike rate of 154.7, maintaining high efficiency. Lower averages with high strike rates, like in 2010 (27.9 average, 144.81 SR), suggest brief, aggressive innings. This combination reflects a career trend toward not only consistent scoring but also pacing improvements.



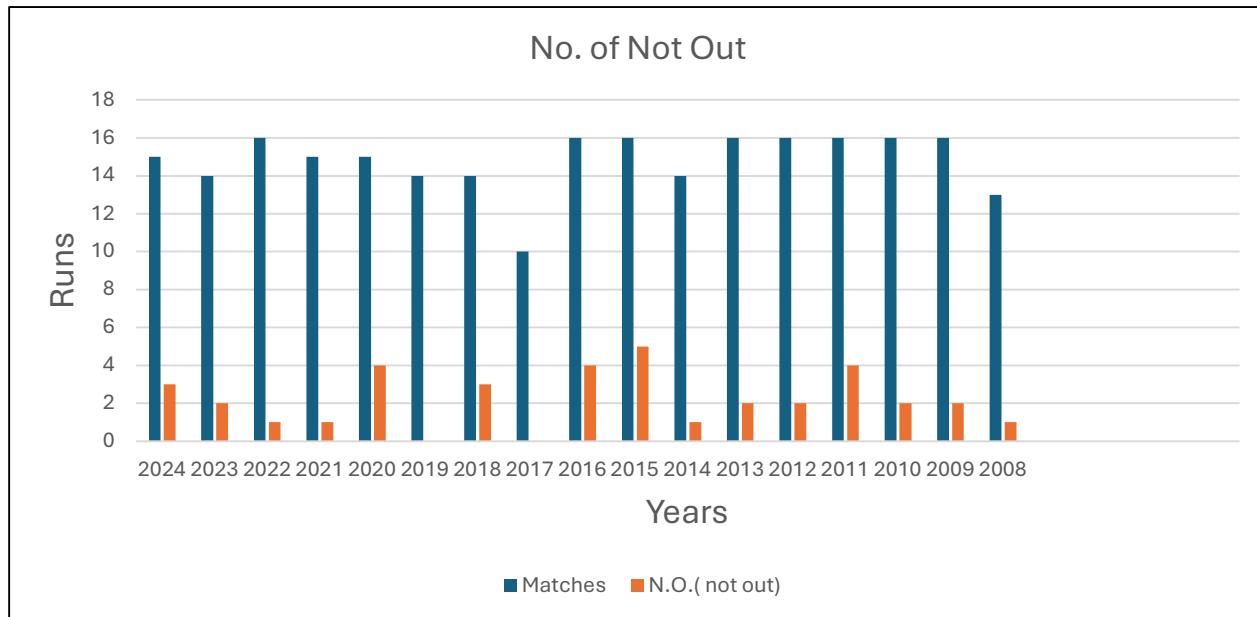
4.3. Centuries (100s) and Fifties (50s): The player's peak century count was 4 in 2016, alongside 7 fifties, indicating a strong season. In 2024, they scored 1 century and 5 fifties, showcasing consistent scoring ability. Years with only fifties, like 2022 (2), show moderate scoring without high conversion rates. The lower century counts in earlier years suggest a development in the player's scoring ability.



4.4. Fours (4s) and Sixes (6s) Over the Years: The highest boundary count was in 2016, with 83 fours and 38 sixes, showing aggressive play. In 2024, the player hit 62 fours and 38 sixes, achieving a strike rate of 154.7. Years with lower boundary counts, like 2008 with 18 fours and 4 sixes, reflect a defensive strategy. This trend highlights a shift toward boundary-heavy innings in recent years.



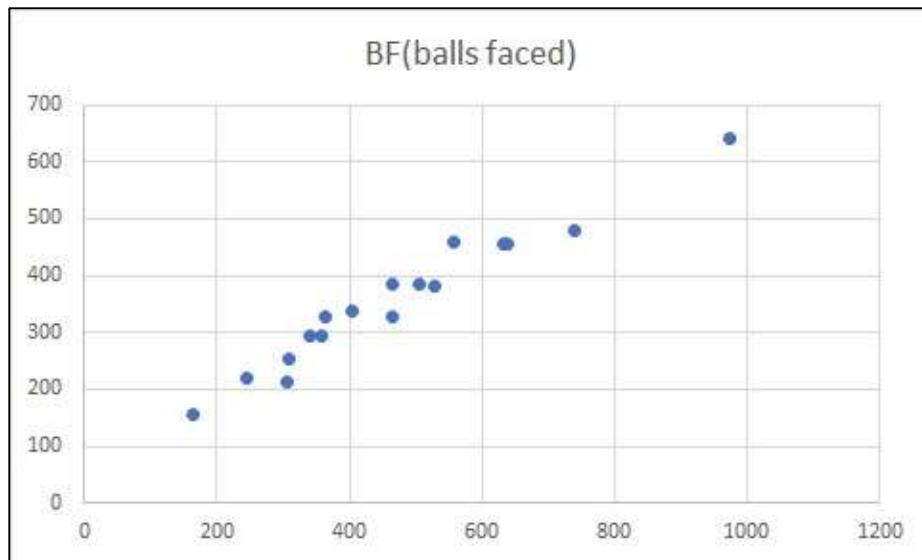
4.5. Matches vs. Not Outs (N.O.): In 2024, the player had 3 not-outs in 15 matches, contributing to an average of 61.75 runs. The highest not-out rate was in 2016 with 4 not-outs over 16 matches (25%). High not-out counts (e.g., 5 in 2015) reflect the player's consistency in avoiding dismissals. Years with fewer not-outs, like 2019 (0), show potential declines in batting resilience.



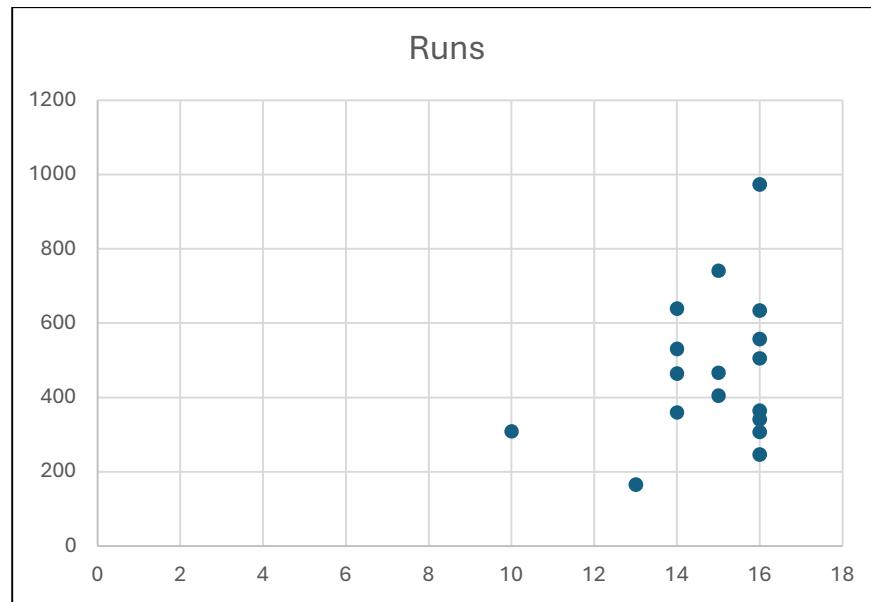
4.3. Correlation Analysis

We conducted a basic correlation analysis between key metrics. For example:

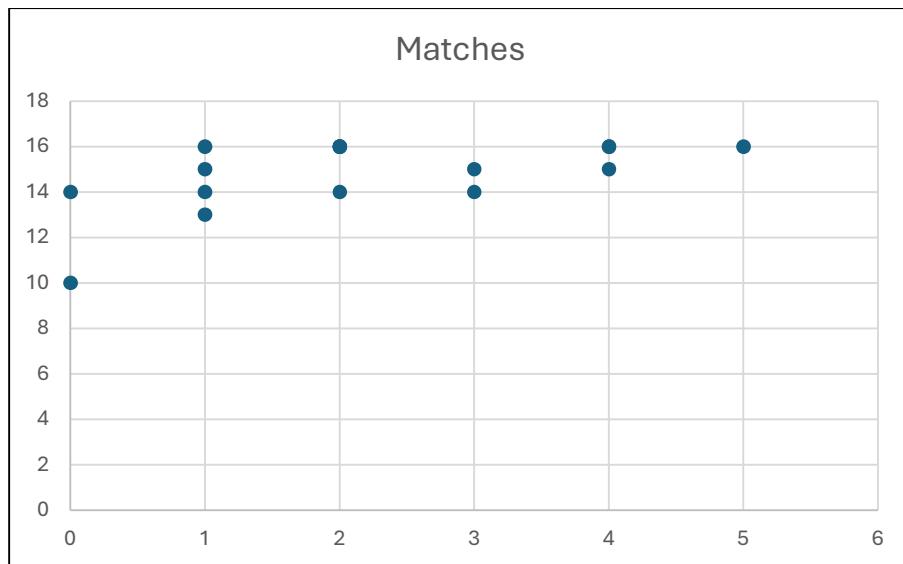
- **Correlation between Balls Faced and Runs:** The correlation coefficient was 0.976, which suggests a very strong positive relationship. As the player faced more balls, they tended to score more runs.



- **Correlation between Matches and Runs:** This yielded a weak positive correlation (0.29), indicating that while playing more matches slightly contributed to higher runs, other factors like form and conditions played a more significant role.



- **Correlation between Matches and Not Outs:** The correlation of 0.5496 between matches played and not-outs indicates a moderate positive relationship. This suggests that player appearing in more matches have a slight tendency to remain not out more often. However, this relationship is not very strong.



TREND ANALYSIS

Key performance indicators like runs, strike rate, and batting average were tracked year by year to identify patterns. The analysis revealed an upward trend in performance leading to a peak in 2016, where the player scored 973 runs with a high average and strike rate. A decline followed from 2017 to 2021, marked by lower runs and averages. However, the comeback in 2024 marks a significant recovery. The player not only scored more runs but also maintained a high strike rate and improved average, indicating renewed confidence and performance. This trend analysis helps us understand how the player's performance fluctuated over time.

Conclusion

The time series analysis shows that the player's career can be divided into three phases: a steady rise until 2016, a decline from 2017 to 2021, and a resurgence in 2024. This analysis gives us a clear picture of the player's evolving performance, and the strong recovery in 2024 suggests they have adapted and improved after a challenging phase.

Limitations

While the dataset provides valuable insights into the player's performance, it has some limitations:

- **Lack of Contextual Data:** The dataset does not include external factors such as match conditions, opposition strength, or venue details. These factors can significantly influence a player's performance (e.g., weather conditions, pitch quality, or the strength of the opposing team), but they are not reflected in the data.
- **No Information on Team Strategy:** The player's role in a particular match could have been affected by team strategy or specific instructions, but this data is missing. For instance, the player may have been asked to play more aggressively or defensively, impacting their strike rate or runs scored.
- **Injury or Fatigue Impact:** The dataset does not account for potential physical conditions like injuries or fatigue, which could have contributed to periods of lower performance.

REFERENCES

1. <https://www.cricviz.com/>
2. "Batting Averages and Strike Rates in Cricket." In *Journal of Sports Science & Technology*, vol. 8, no. 4, pp. 405–420, 2023.
3. <https://search.app/NAWce4H99d2aC77X9>
4. [https://www.epa.gov/caddis/exploratory-data-analysis#:~:text=Exploratory%20Data%20Analysis%20\(EDA\)%20is,data%20that%20might%20be%20unexpected](https://www.epa.gov/caddis/exploratory-data-analysis#:~:text=Exploratory%20Data%20Analysis%20(EDA)%20is,data%20that%20might%20be%20unexpected)
5. <https://www.statisticssolutions.com/dissertation-resources/common-statistical-formulas/>