

# IPL Data Analysis

**IT-495 Exploratory Data Analysis**  
**Project Report**  
**Semester-II**  
**Instructor : Gopinath Panda**

**Group : 9**

**Rohan Shah** (202218002)



**Karan Parashar** (202218004)



**Nauman Shaikh** (202218010)



**Nisarg Shah** (202218034)

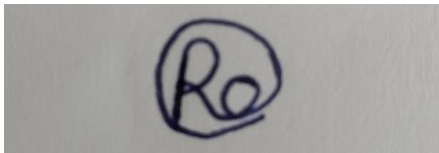


---

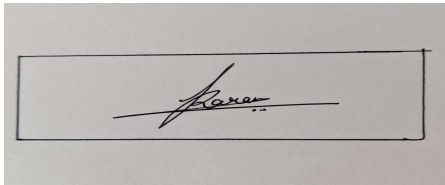
## Declaration:

We hereby declare that this project report entitled "IPL Data Analysis" is a presentation of our own research work and has not been submitted to any other university or college for any award. Wherever contributions of others are involved, every effort is made to indicate that clearly with due reference to the literature and acknowledgment of collaborative research and discussions.

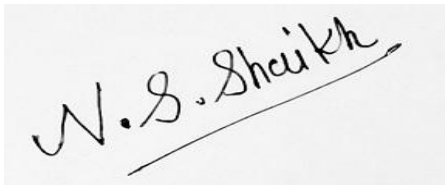
Sign of group members



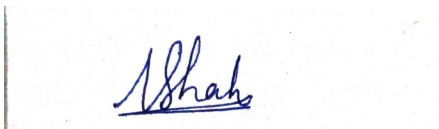
(202218002 - Rohan Shah)



(202218004 - Karan Parashar)



(202218010 - Nauman Shaikh)



(202218034 - Nisarg Shah)

---

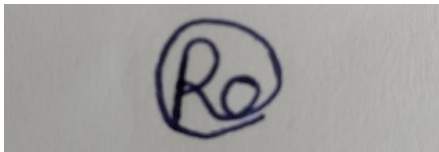
## Certificate

Rohan Shah(202218002), Karan Parashar (202218004), Nauman Shaikh(202218010), Nisarg Shah(202218034), have submitted their project report entitled "IPL Data Analysis". This project has been guided by Dr. Gopinath Panda, DAICT, Gandhinagar. This project has not been submitted to any other institution for the award of any degree or diploma.

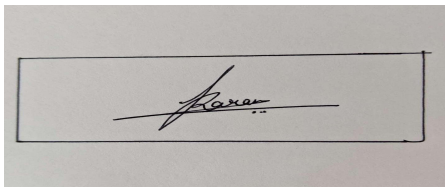
Dr. Gopinath Panda

DAICT, Gandhinagar

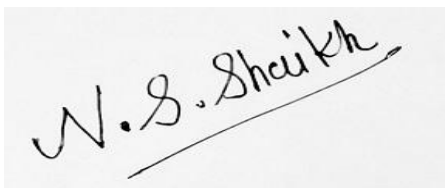
Sign of group members

A handwritten signature in blue ink, consisting of the letters 'R' and 'o' enclosed in a circle.

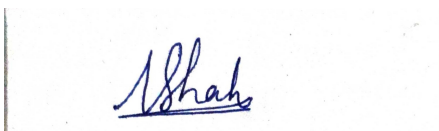
(202218002 - Rohan Shah)

A handwritten signature in black ink, enclosed within a rectangular box. The signature appears to be 'Karan'.

(202218004 - Karan Parashar)

A handwritten signature in black ink, reading 'N.S. Shaikh' with a long horizontal stroke extending from the end.

(202218010 - Nauman Shaikh)

A handwritten signature in blue ink, reading 'Nisarg' with a horizontal line underneath.

(202218034 - Nisarg Shah)

---

## Table of Contents

Declaration.....	2
Certificate.....	3
Table of Contents.....	4
List of figures.....	5
<b>Introduction.....</b>	<b>6</b>
<b>Dataset Description.....</b>	<b>6</b>
<b>Data Reading.....</b>	<b>7</b>
<b>Exploratory Data Analysis:.....</b>	<b>8</b>
<b>Machine Learning.....</b>	<b>28</b>
<b>References.....</b>	<b>34</b>
<b>CV.....</b>	<b>35</b>

---

## List of figures

- [Fig 1.1: Heatmap for null values](#)
- [Fig 1.2: Runs vs. Over](#)
- [Fig 1.3: Total runs vs ball of over](#)
- [Fig 1.4: Wickets per over](#)
- [Fig 1.5: Total number of 6s per season](#)
- [Fig 1.6: Total number of 4s per season](#)
- [Fig 1.7: Total number of boundaries per season](#)
- [Fig 1.8: Over wise run comparison](#)
- [Fig 1.9: Type of Wicket](#)
- [Fig 1.10: Highest number of sixes by individual](#)
- [Fig 1.11: Highest number of fours by individual](#)
- [Fig 1.12: Highest number of ball faced by individual](#)
- [Fig 1.13: Highest runs scored by individual in single match](#)
- [Fig 1.14: Highest number of half centuries by individual](#)
- [Fig 1.15: Highest number of centuries by individual](#)
- [Fig 1.16: Highest run scorer in IPL](#)
- [Fig 1.17: Highest strike rate](#)
- [Fig 1.18: Highest wicket taker](#)
- [Fig 1.19: Most ball bowled by bowler](#)
- [Fig 1.20: Best Spell \(Wickets taken in single match\)](#)
- [Fig 1.21: Highest 4 wicket howl by bowler](#)
- [Fig 1.22: Google search trend for IPL](#)
- [Fig 2.1: Run vs Wickets Scatter plot](#)
- [Fig 2.2: Elbow method for optimal number of clusters](#)
- [Fig 2.3: for clusters  \$k = 3\$ , Runs vs Wicket](#)
- [Fig 2.4 : Regression line for V Kohli to predict runs](#)
- [Fig 2.5 : Regression line for MS Dhoni to predict runs](#)
- [Fig 2.6 : Regression line for V Kohli to predict runs](#)

---

## Introduction

In this project, we have done an analysis of IPL matches between 2008-2023 using exploratory data analysis techniques. The programming language used is Python 3. The modules used are Numpy, Pandas, for data analysis and manipulation. For visualization, we have used modules such as Matplotlib, Seaborn, Plotly. We have used the Pytrends module for analysis of google trends about IPL. For predictive analysis, we have used the scikit-learn module for clustering, classification and regression.

## Dataset Description

The data is in .csv (Excel Comma Separated Values) format, containing ball by ball data of IPL matches between 2008 and 2023. The data was originally obtained from [https://cricsheet.org/downloads/ipl\\_csv2.zip](https://cricsheet.org/downloads/ipl_csv2.zip), and modified to add some extra features.

The modified dataset can be found here:

<https://drive.google.com/file/d/1esH6mZ3ez4eirh5qjxNiS9-Wp2l1TUjN/view?usp=sharing>

The dataset contains 234759 rows and 25 columns/attributes. The important features are described as follows:

Attribute	Description
id	Match Id
inning	innings of the match that was being played
over	Over of the match
ball	Ball (1 to 6)
batsman	Batsman on strike
non_striker	Non-striker batsman
bowler	Current Bowler
batsman_runs	Runs scored by batsman in current ball
extra_runs	Extra runs in the current ball (wide, no ball etc.)

---

total_runs	Total of batsman_runs and extra_runs
non_boundary	Whether boundary was not scored
is_wicket	Whether wicket was taken in this ball
dismissal_kind	If wicket was taken, type of dismissal
player_dismissed	If wicket was taken, player which was dismissed
extras_type	Type of extra runs (wide, no ball etc.)
batting_team	Batting Team
bowling_team	Bowling Team

### Attribute Types:

- **Nominal** - id, batsman, non\_striker, bowler, dismissal\_kind, player\_dismissed, fielder, extras\_type, batting\_team, bowling\_team
- **Ordinal** - over, ball, batsman\_runs, total\_runs, extra\_runs
- **Binary** - inning, is\_wicket

### NA and Null Type values:

dismissal\_kind, player\_dismissed, fielder, extras\_type contain NAN type values. It contains definite value only when there is a wicket (dismissal) or extras involved.

Percentage of Total Missing Values is 35.39 %. The Feature dismissal\_kind has 223136 missing values. The Feature player\_dismissed has 223136 missing values.

## Data Reading

The data in .csv format is loaded into a *pandas.DataFrame* object, using the Panda's *read\_csv* method as follows:

```
data = pd.read_csv("/content/drive/MyDrive/ipl_all_matches.csv")
```

---

## Exploratory Data Analysis:

- Heatmap for null values:

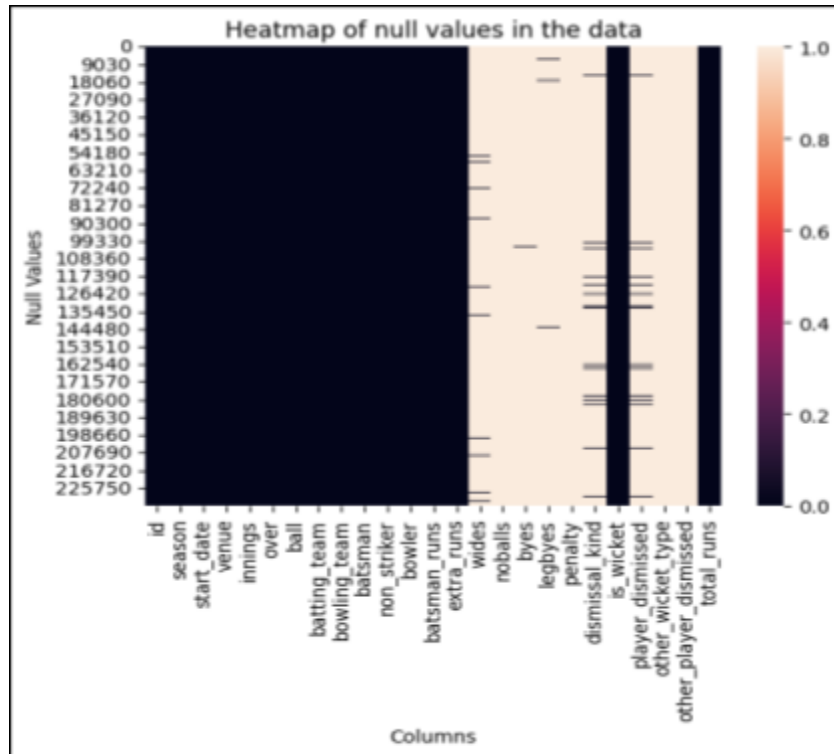


Fig 1.1 : Heatmap for null values

The heatmap shows that dismissal\_kind, player\_dismissed, fielder, extras\_type contain many NAN values. This is because the value is filled only if dismissal or extras are involved.



---

- **Finding total runs scored in each over:**

From our data we can say that in the 1st over of every match, a combined 12107 has been scored. And for other overs, graph is shown below:

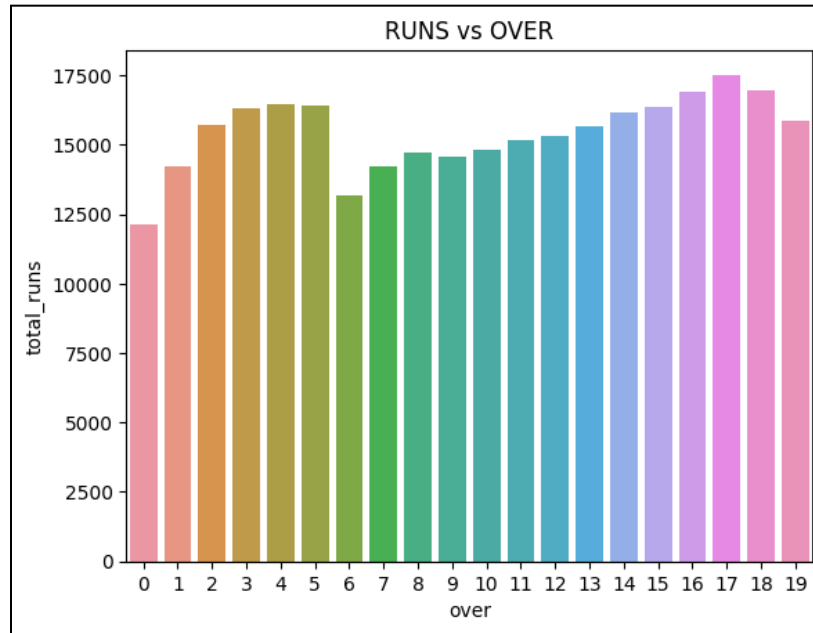


Fig 1.2: Runs vs. Over

Also karl pearson's correlation coefficient for given data is : **0.5781**

- 
- **Finding Runs scored on each ball :** We can say that from our analysis that on the first ball of every over for every match a combined 48594 runs have been scored. We can also see the runs for other balls as shown in the graph below.

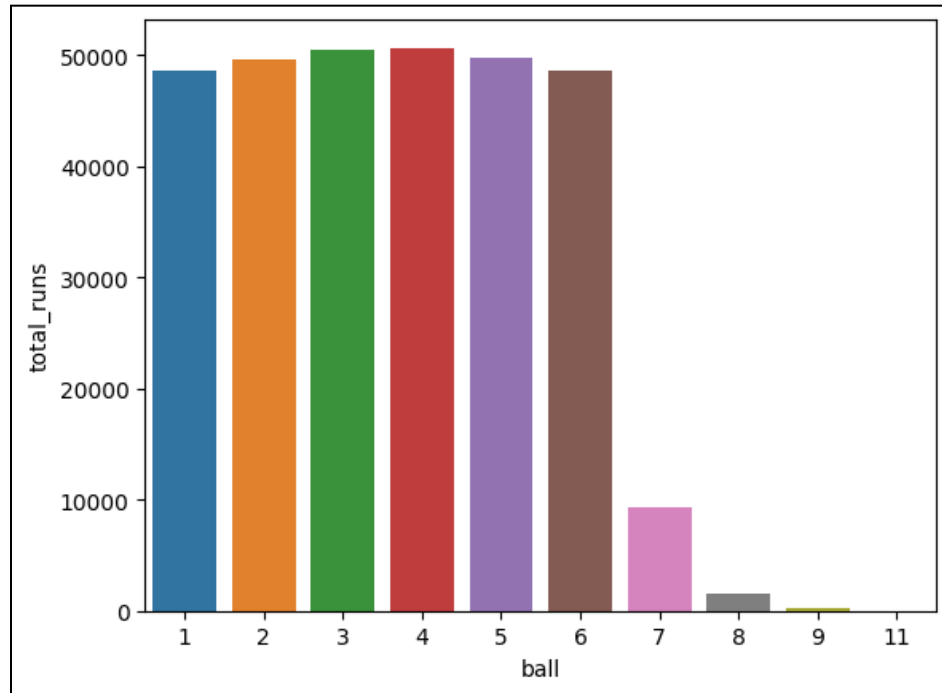


Fig 1.3: Total runs vs ball of over

Correlation coefficient is : **-0.8659**

- 
- **Finding wickets taken in each over:** From this we conclude that most wickets have been taken in the 20th over, which is 1187. The reason behind is because it is the last over of the match, resulting in the batsman taking risk and hitting big hits, causing losing his/her wicket.

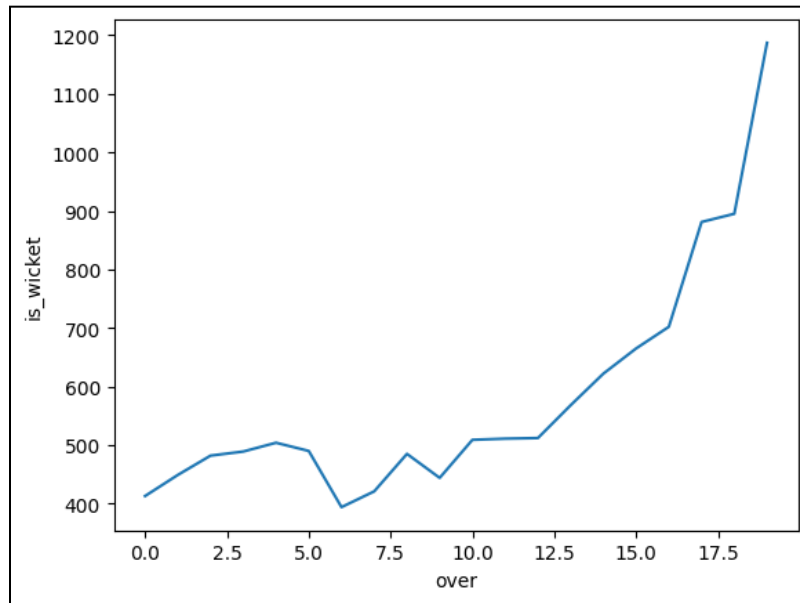


Fig 1.4 : Wickets per over.

Correlation between over and no. wicket is: **0.7961**

---

**(1) Analysis of 4's and 6's season wise:**

From our analysis we conclude that 2018 was the season with the highest number of sixes which is 853. From another season, no of sixes is shown in the graph below.

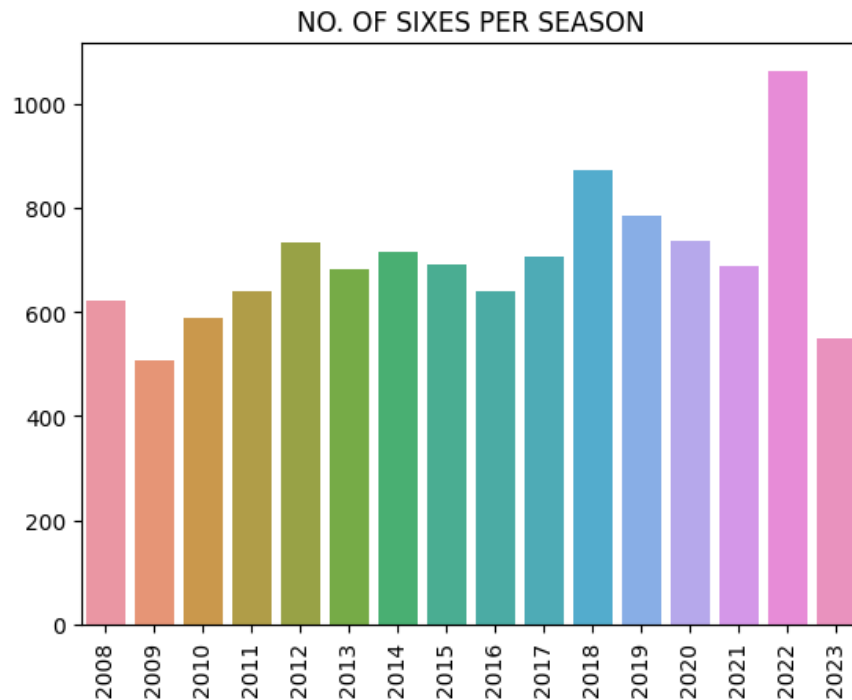


Fig 1.5: Total number of 6s per season

Similarly we conclude that the most number of fours has been scored in season 2013 with 2025 four.

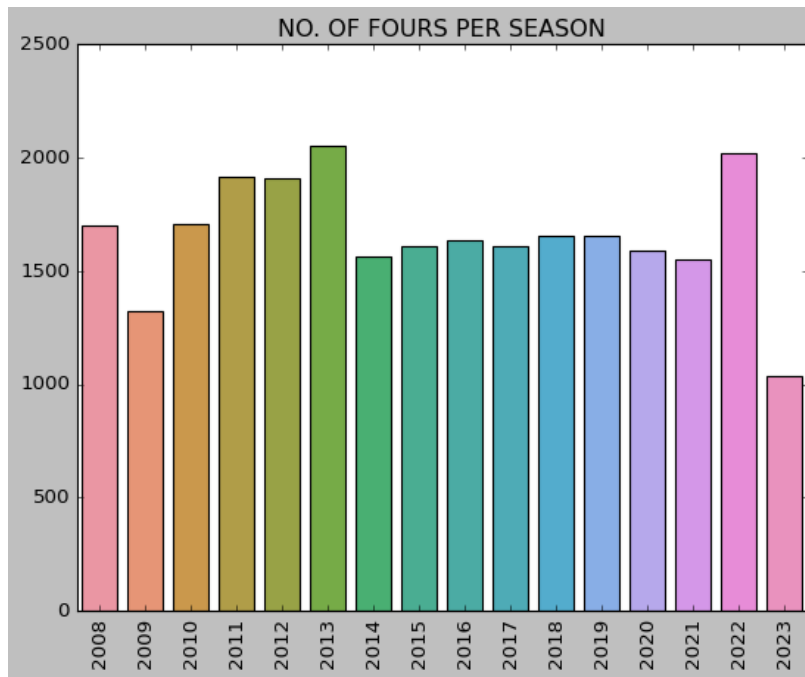


Fig 1.6: Total number of 4s per season

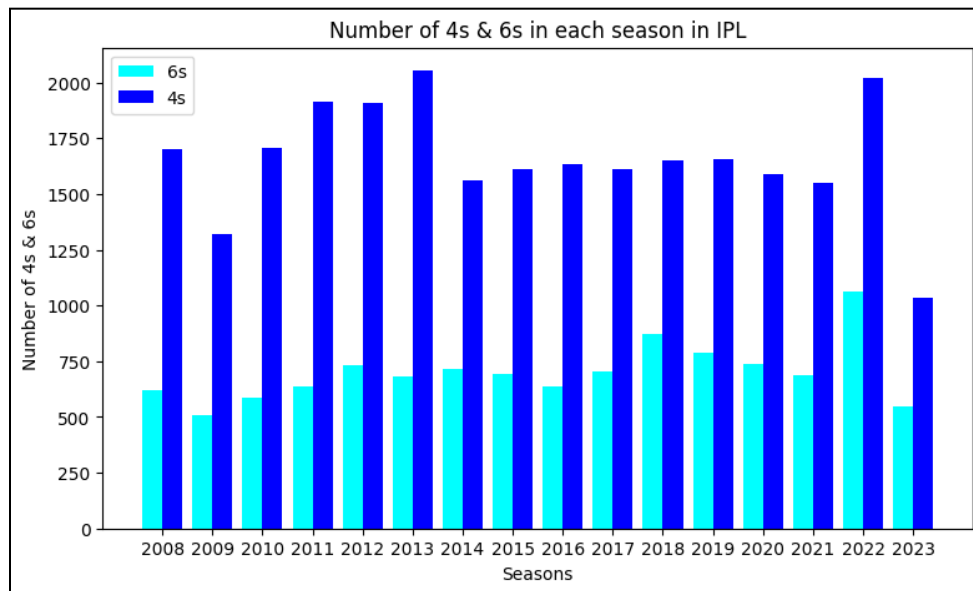


Fig 1.7: Total number of boundaries per season

- Graph below is showing the percentage of runs scored in particular overs. We can see that most runs have been scored in the 17th over.

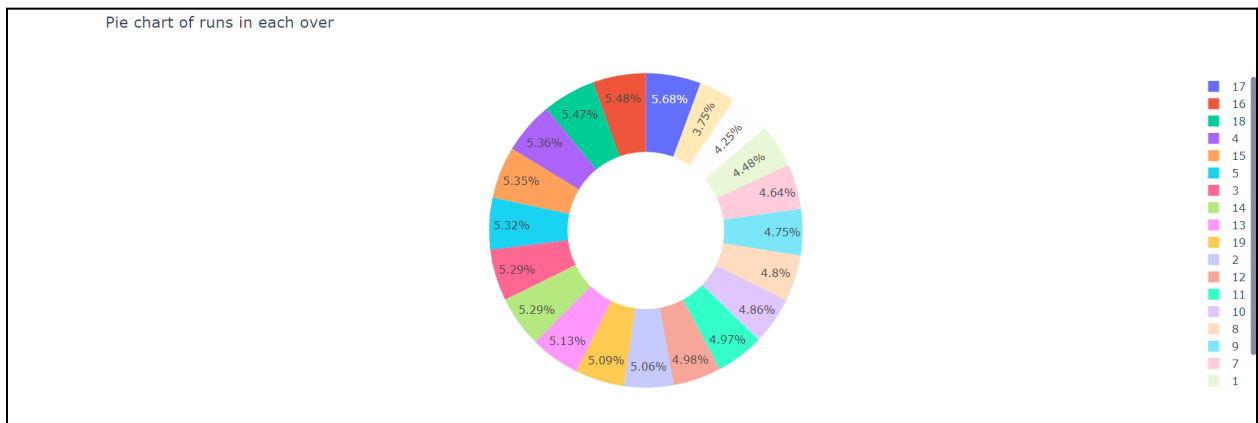


Fig 1.8: Over wise runs comparison

- Below Graph is showing percentage of wickets fallen and its cause. We conclude that the reason for the majority of wickets to fall is 'Catch out', followed by 'Bowled'.

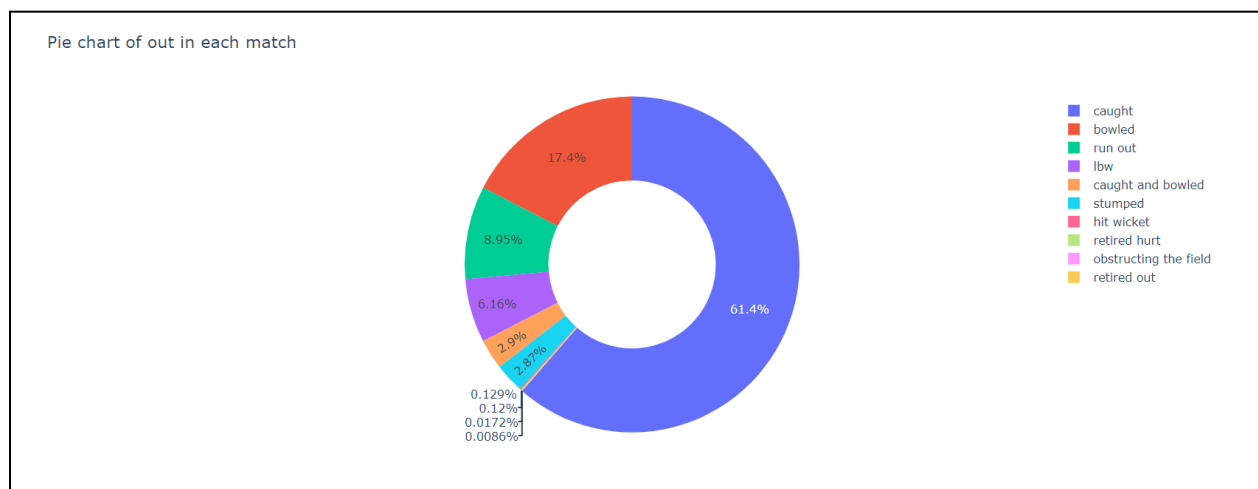


Fig 1.9: Type of wicket

---

- **Batsman Stats Analysis:**

Highest number of sixes scored by a player: Chris Gayle - 349 (6's)

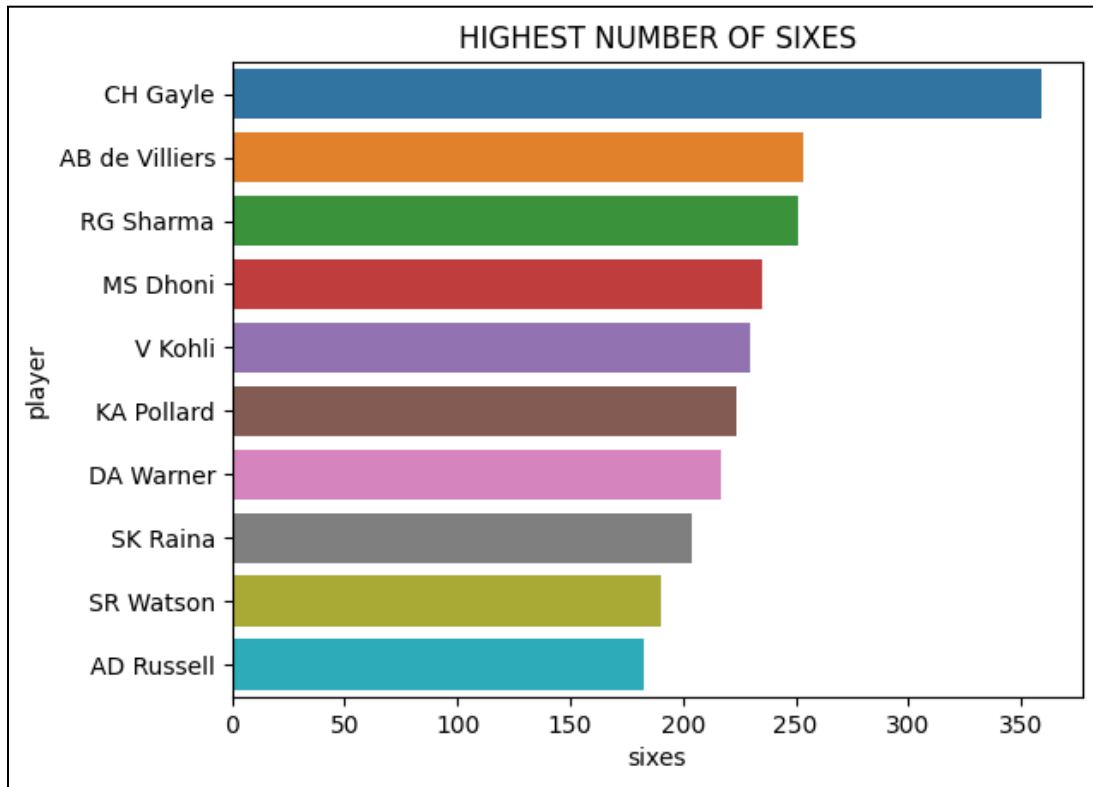


Fig 1.10: Highest Number of 6s by individual

---

Similarly highest number of fours scored by player: Shikhar Dhawan - 730

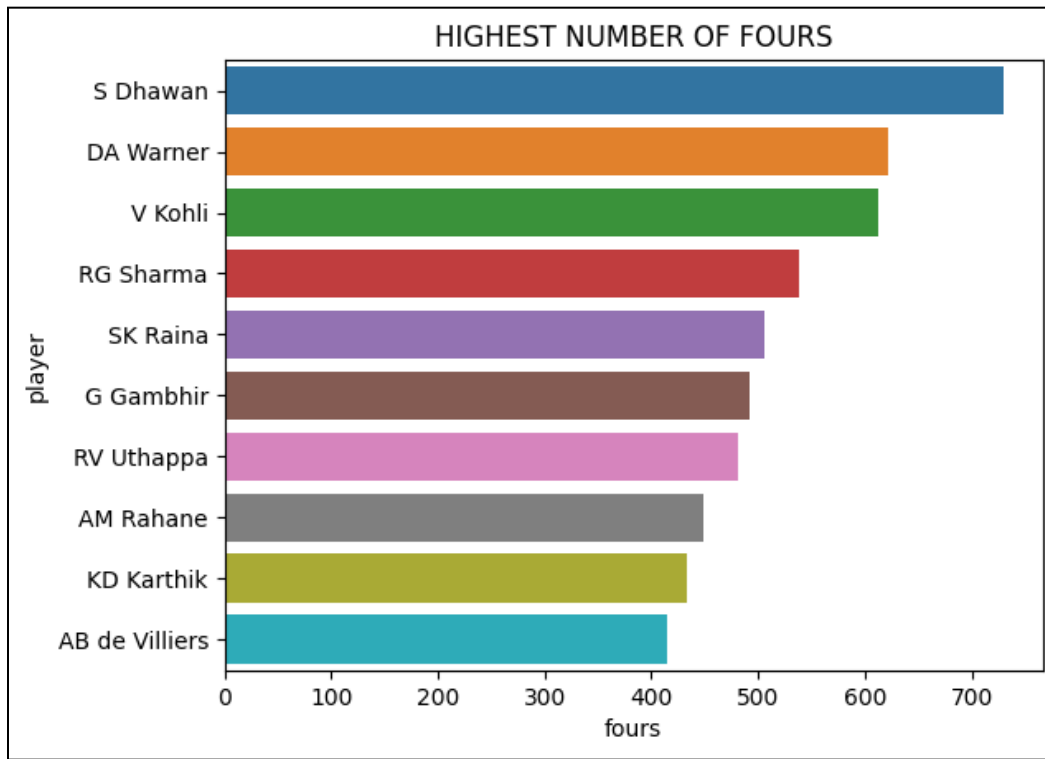


Fig 1.11: Highest number of 4s by individual



---

Similarly Highest ball faced by a player : Virat Kohli - 5301

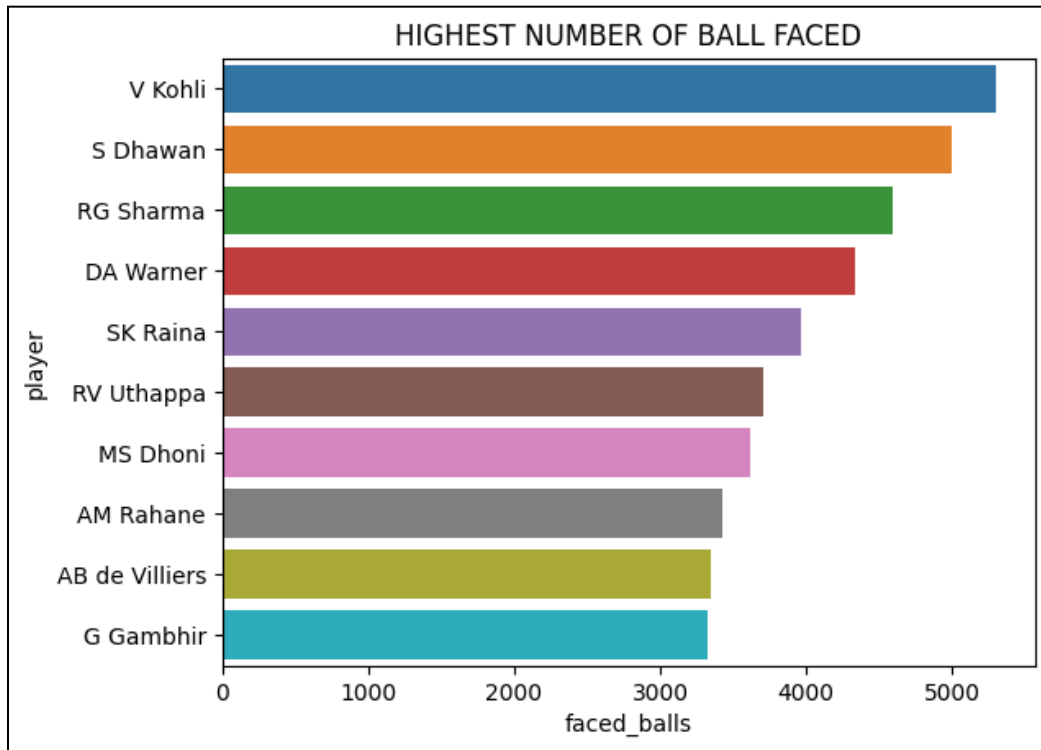


Fig 1.12: Highest number of ball faced by individual

---

Highest run scored by individual player: Chris Gayle - 175

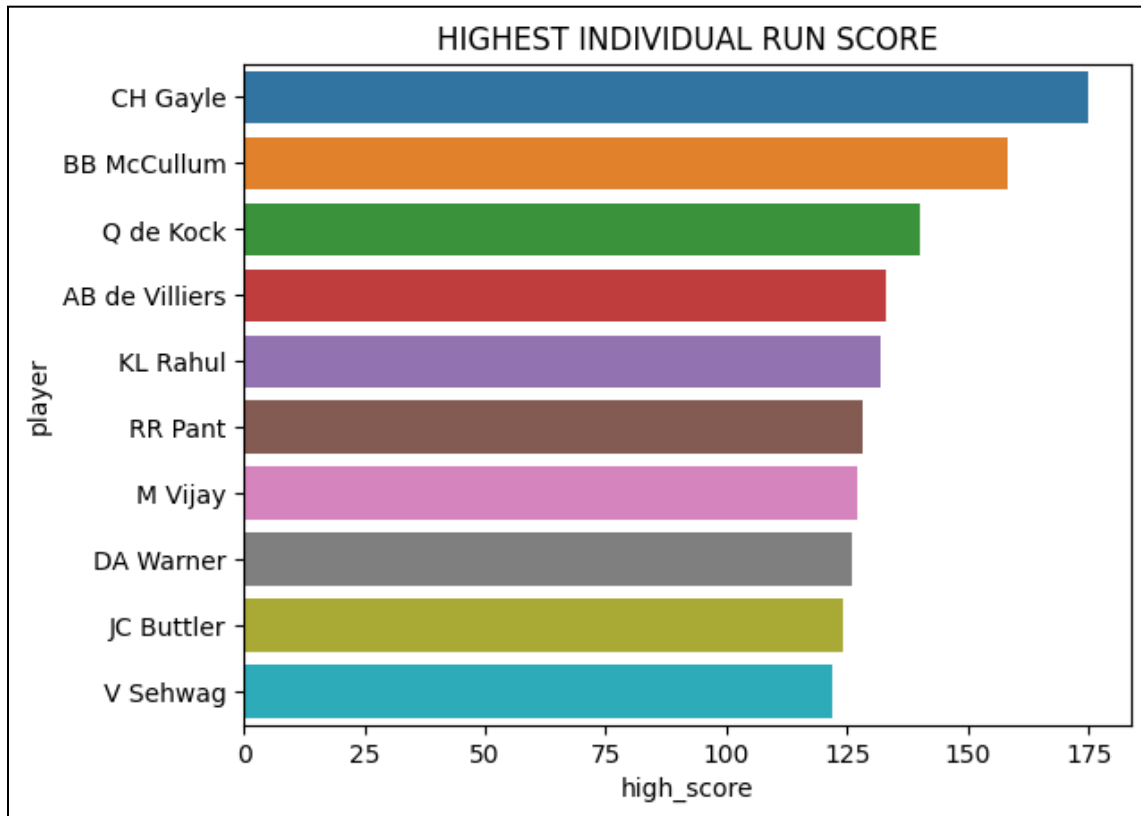


Fig 1.13 Highest runs scored by individual in single match

---

Most number of half centuries scored by a individual player: David Warner - 59

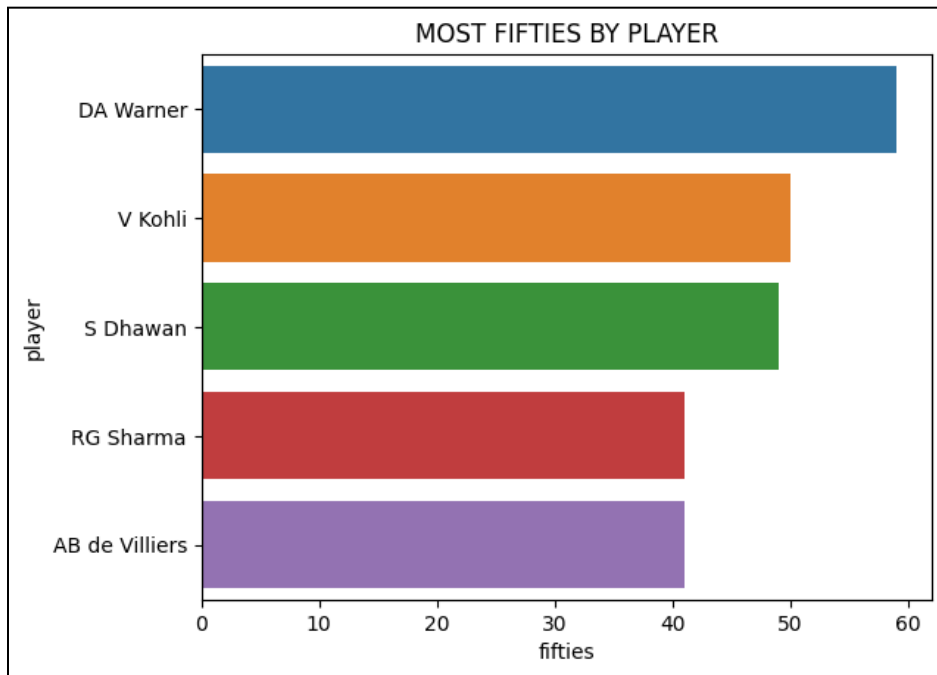


Fig 1.14: Most half centuries by individual

---

Most number of Centuries scored by individual player: Chris Gayle - 6

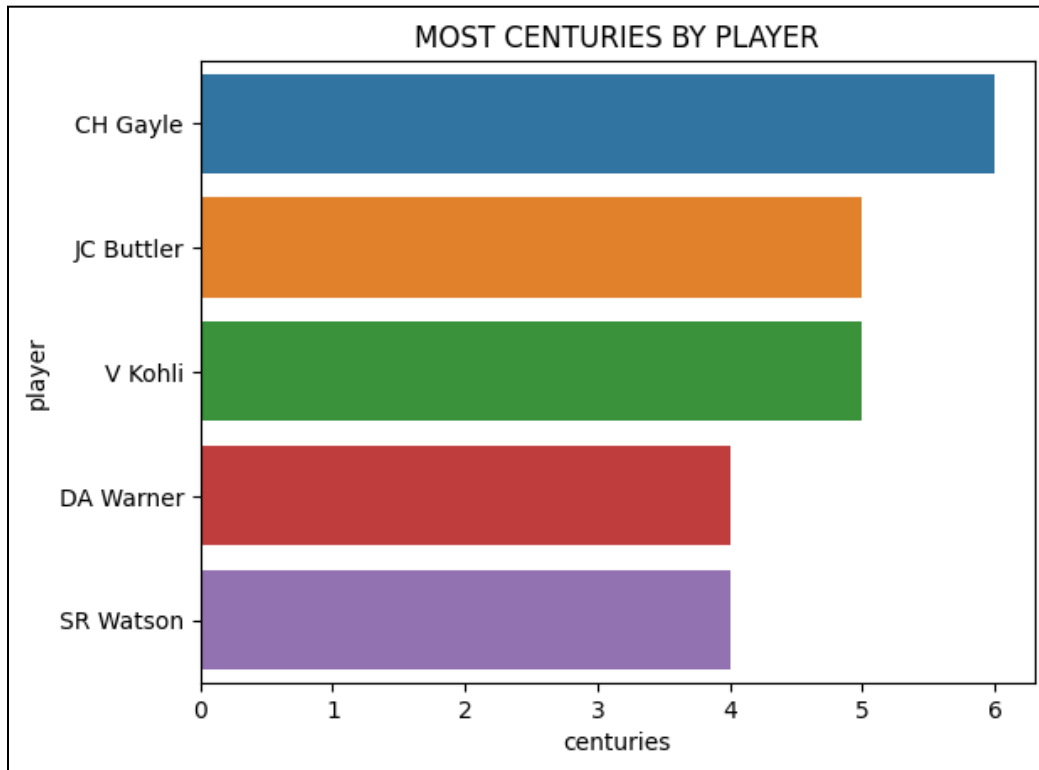


Fig 1.15: Most number of centuries by individual

After that we computed 2 new columns named, batting average and strike rate.

Strike Rate = Total Runs scored \* 100 / Total Balls Faced

Batting AVG = Total Runs scored / total dismissal

- Top run scorer in IPL: Virat Kohli - 6967

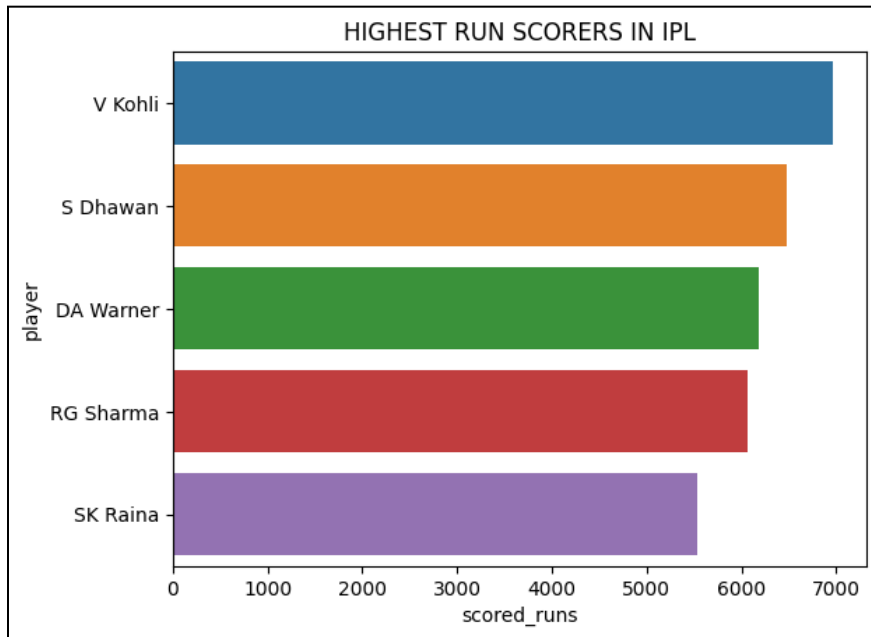


Fig 1.16: Highest Run scorer in IPL

---

Similarly calculated, Highest Strike rate: PN Mankad - 398

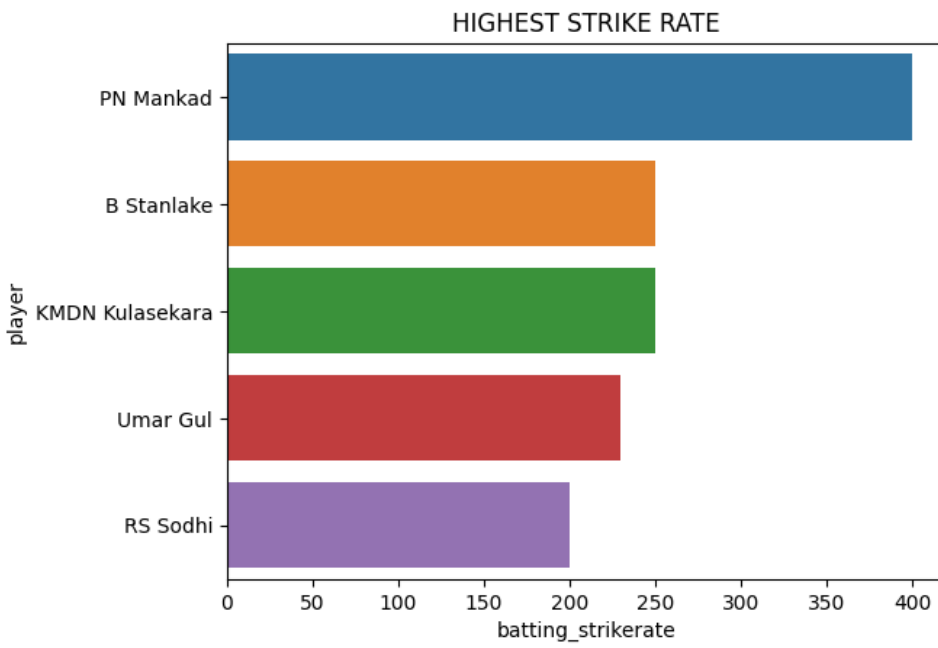


Fig 1.17: Highest batting strike rate

- 
- **Bowler Analysis**
  - **Top bowlers by wickets**

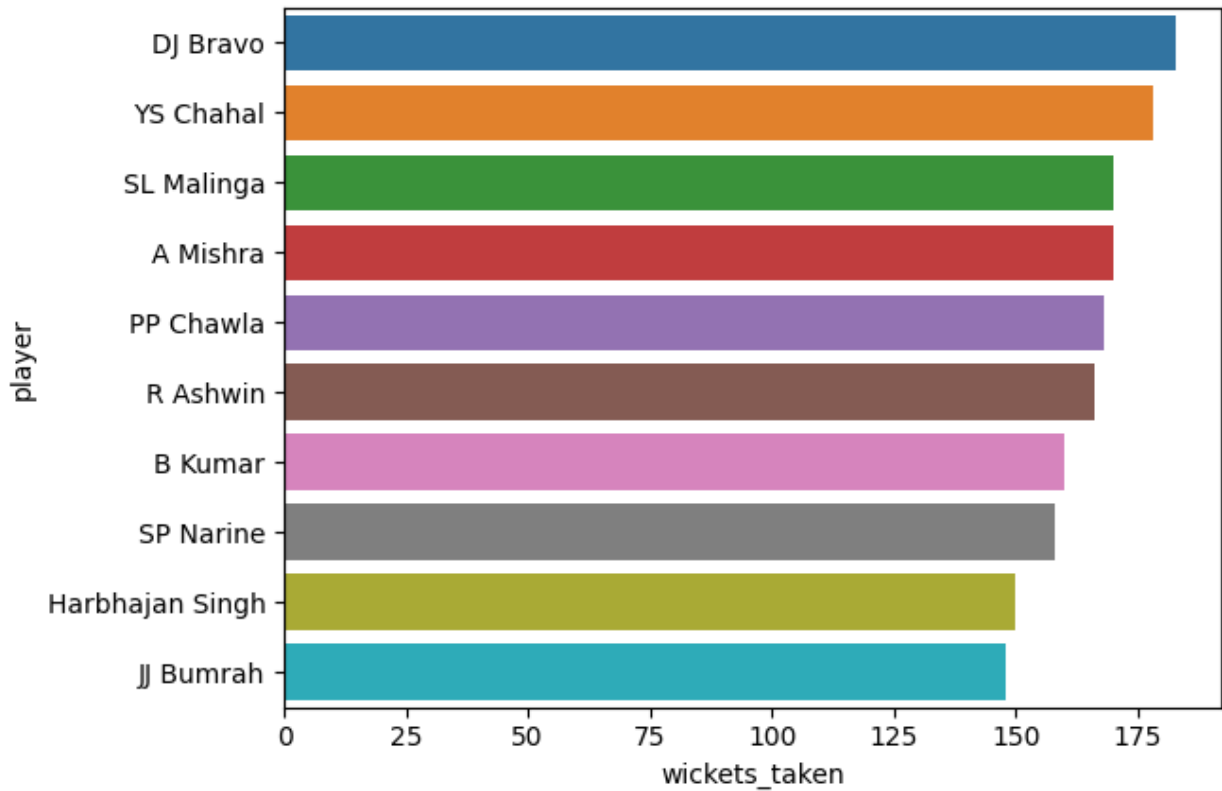


Fig 1.18: Highest wicket taker bowler

Here, we can see that DJ Bravo is the top bowler in terms of taking wickets, followed by YS Chahal and SL Malinga.

---

- **Top bowlers by balls bowled:**

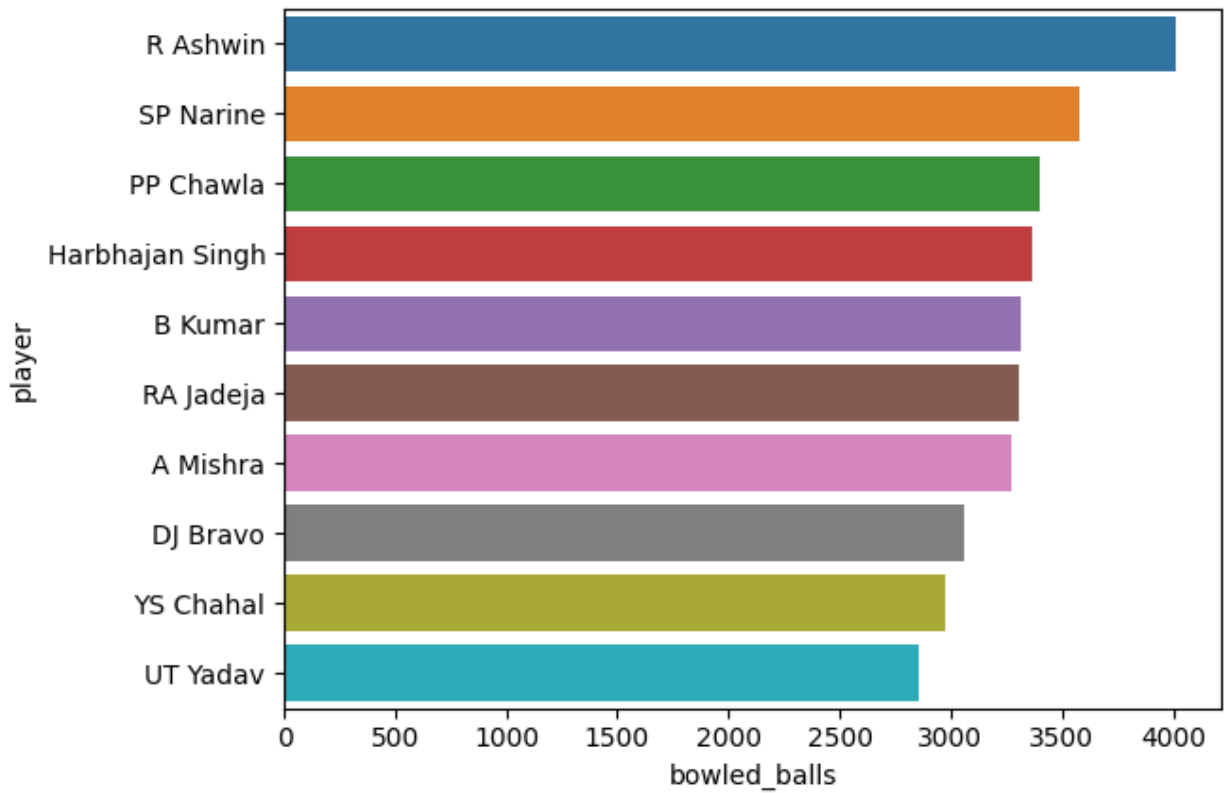


Fig. 1.19: Highest Balls bowled by bowler

Spinner R Ashwin has bowled the most balls (around 4016), followed by SP Narine and PP Chawla. The top 8 bowlers here have bowled more than 3000 balls.



---

- **Top Bowlers by Best Spell:**

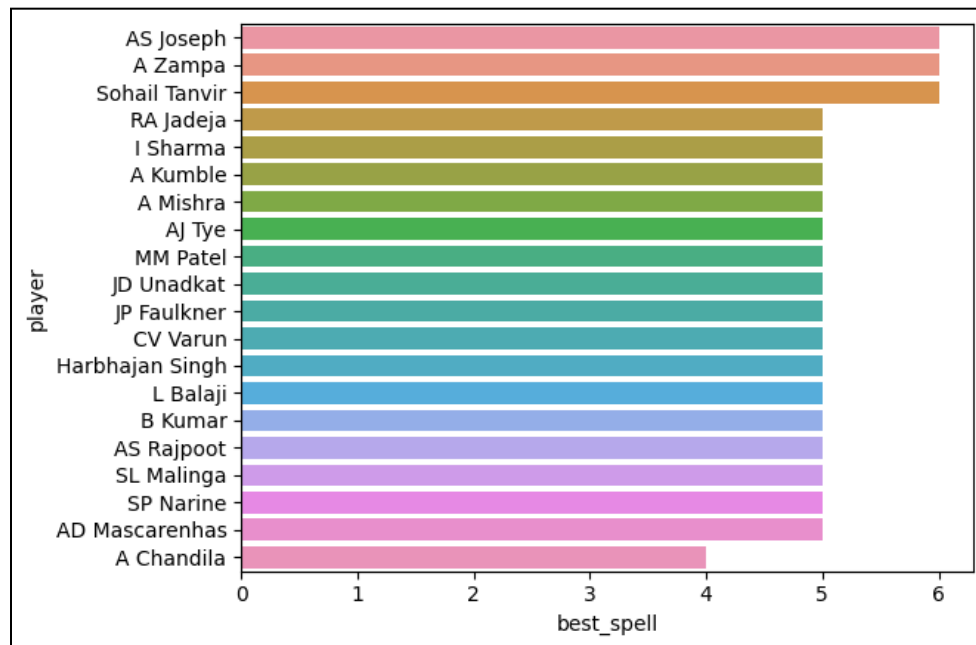


Fig 1.20: Best Spell (Wickets taken in single match)

This plot shows us the best performance of bowlers in each match sorted by number of wickets in the spell. The best performing bowlers in this case are AS Joseph, A Zampa, and Sohail Tanvir.

---

- **Bowlers taking more than 4 wickets**

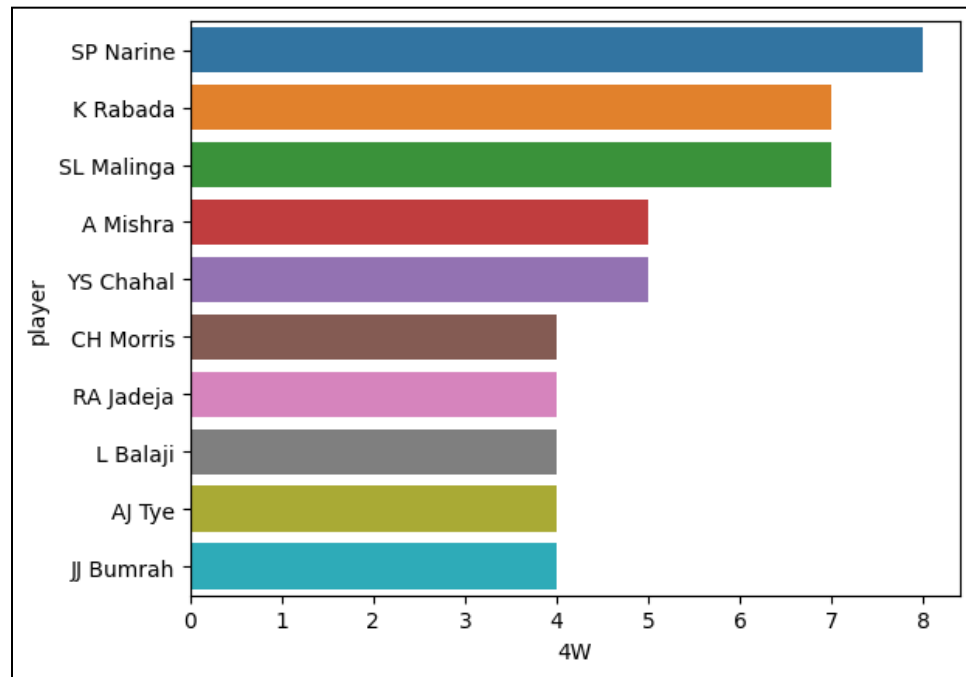


Fig 1.21: Highest 4 wicket howl by bowler

This plot shows us the bowler, and the number of matches in which he has taken more than 4 wickets. The best performing bowlers in this case is SP Narine, followed by K Rabada and L Malinga.

---

- **PyTrends analysis**

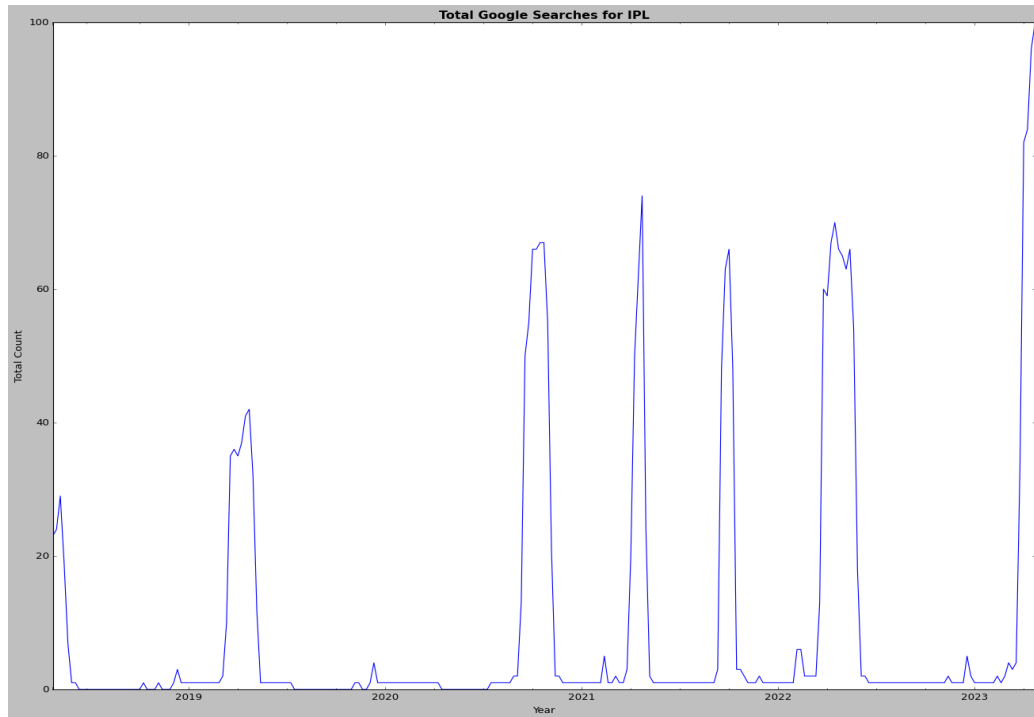


Fig 1.22: Google search trend

This plot shows the trend of google searches for IPL over time. As we can see, the searches were low in 2019, but have increased significantly nearing 2023.

---

## Machine Learning

We have tried to implement machine learning techniques such as classification, regression and clustering.

Consider the below scatter of runs vs. wickets for all players. We aim to apply k-means clustering to group together different types of players.

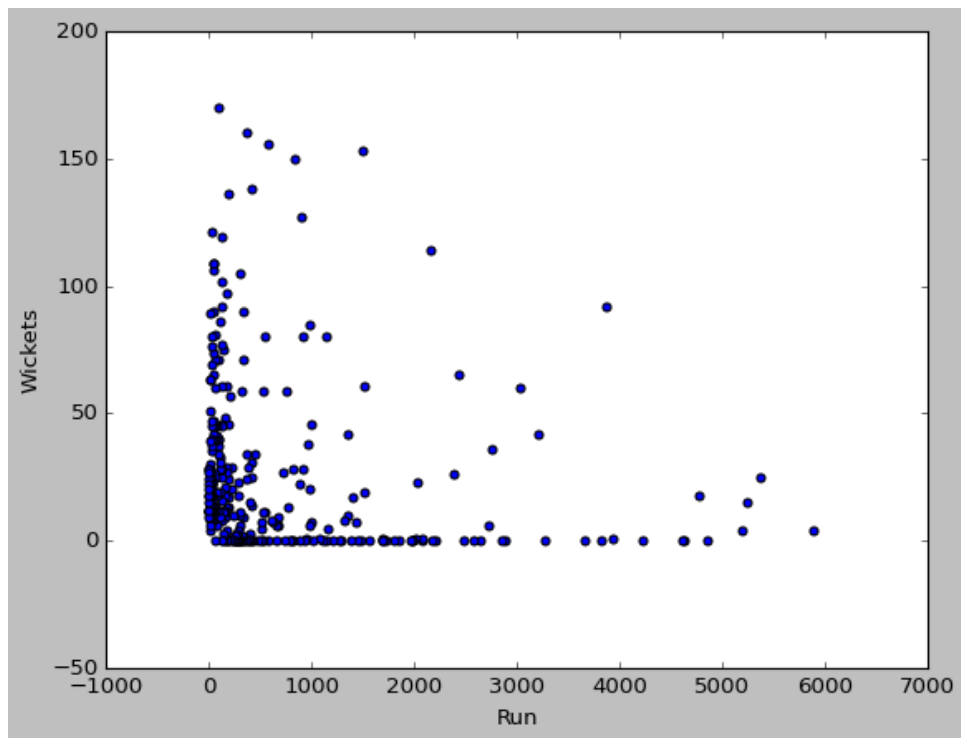


Fig 2.1: Run vs Wickets Scatter plot

To decide the number of clusters, we have plotted an elbow plot as follows:

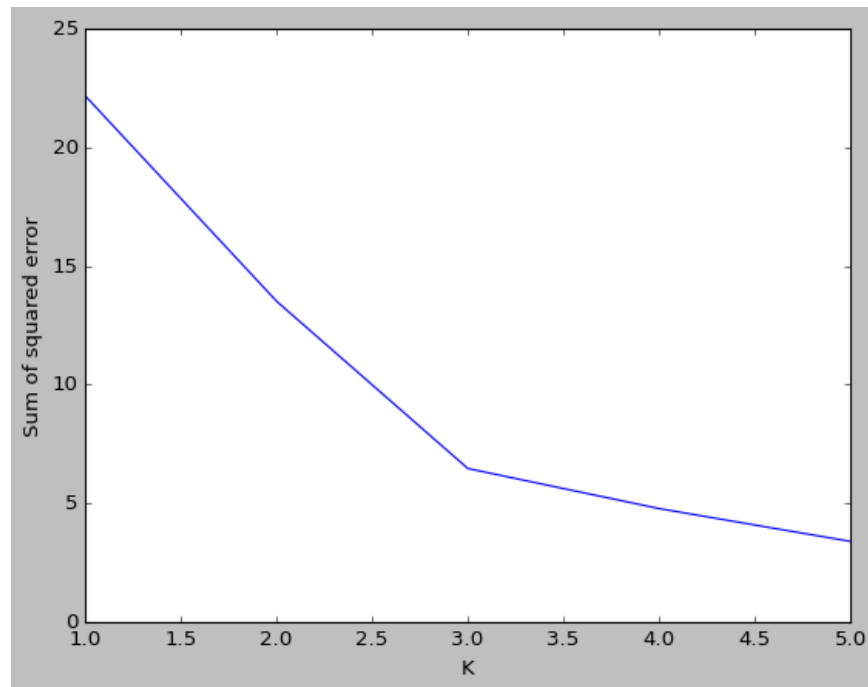


Fig 2.2: Elbow method for optimal number of clusters

Since the error does not reduce significantly after  $k = 3$ , the optimal number of clusters is 3. Below is the result of clustering, based on 3 clusters:

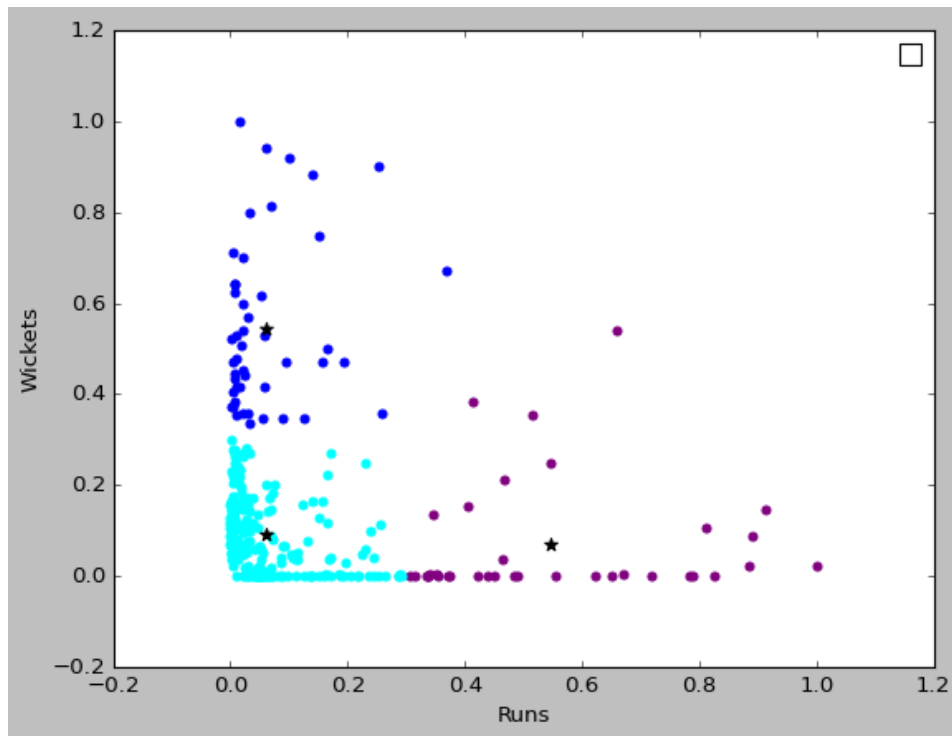


Fig 2.3: for clusters  $k = 3$ , Runs vs Wicket

The best guess is that the clusters are batsman, bowlers, and all-rounders. (Note that the scale for runs and wickets is changed because it was normalized before running the algorithm.)

Based on the identified clusters, we aim to make a classification model that can predict the type of player. The features used to classify the players are 'scored\_runs', 'wickets\_taken', 'batting\_innings', 'bowling\_innings', 'high\_score', 'best\_spell', 'batting\_avg', 'bowling\_avg'. We used a KNN-classifier with the number of neighbors = 5, with the default *minkowski* metric.

---

Following is the confusion matrix:

<b>Actual Predicted</b>	<b>Class 1</b>	<b>Class 2</b>	<b>Class 3</b>
<b>Class 1</b>	9	0	0
<b>Class 2</b>	0	23	3
<b>Class 3</b>	0	1	31

**Interpretation:**

The first row, [9, 0, 0], represents the actual class 1. The classifier correctly predicted 9 instances of this class and did not mistakenly predict any other class as class 1.

The second row, [0, 23, 3], represents the actual class 2. The classifier correctly predicted 23 instances of this class and incorrectly predicted 3 classes as class 3.

The third row, [0, 1, 31], represents the actual class 3. The classifier correctly predicted 31 instances of this class, but incorrectly predicted 1 instance as class 2.

Results of Classifier are as follows:

<b>Class</b>	<b>Accuracy</b>	<b>Precision</b>	<b>Recall</b>	<b>F1-Score</b>
<b>1</b>	0.9718	1	0.8889	0.9412
<b>2</b>	0.9718	0.9737	0.9737	0.9737
<b>3</b>	0.9718	0.9167	1	0.9565

---

- **Linear Regression:**

- **For 1 input feature:**

Here we are predicting the run score by any player, given past performance in IPL.

We are taking 'year' as an independent variable and run\_scored as a dependent or target variable and trying to predict the score of the player using linear regression.

- **For Virat Kohli:**

- Regression equation is:  $11.0143 \cdot x + (-21751.51)$
- Predicted Outcome for 2023 = 530.39 ~ 530 runs.

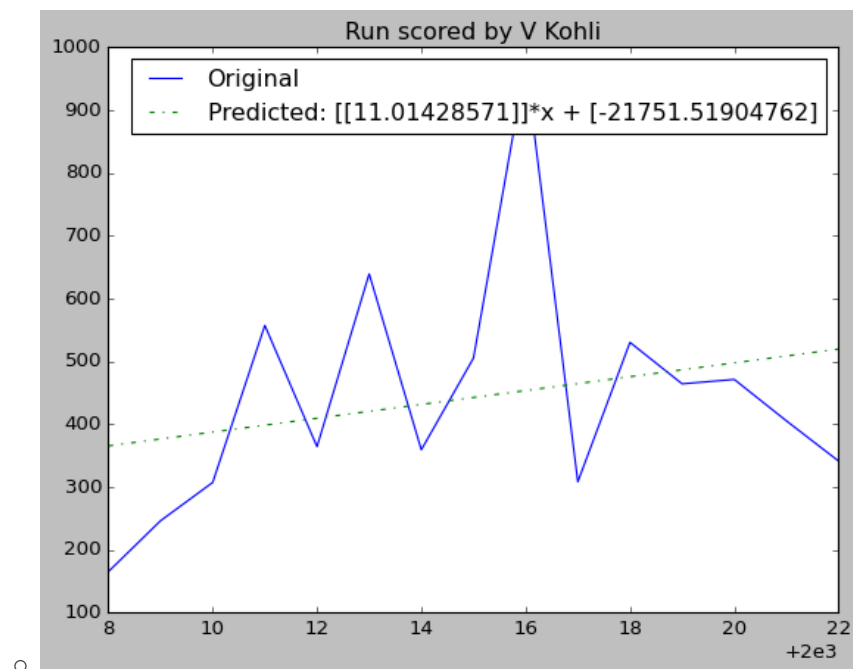


Fig 2.4 : Regression line for V Kohli to predict runs

- **For MS Dhoni:**

- Regression equation is:  $(-10.925) \cdot x + 22345.77$
- Predicted Outcome for 2023: 244.466 ~ 245 runs



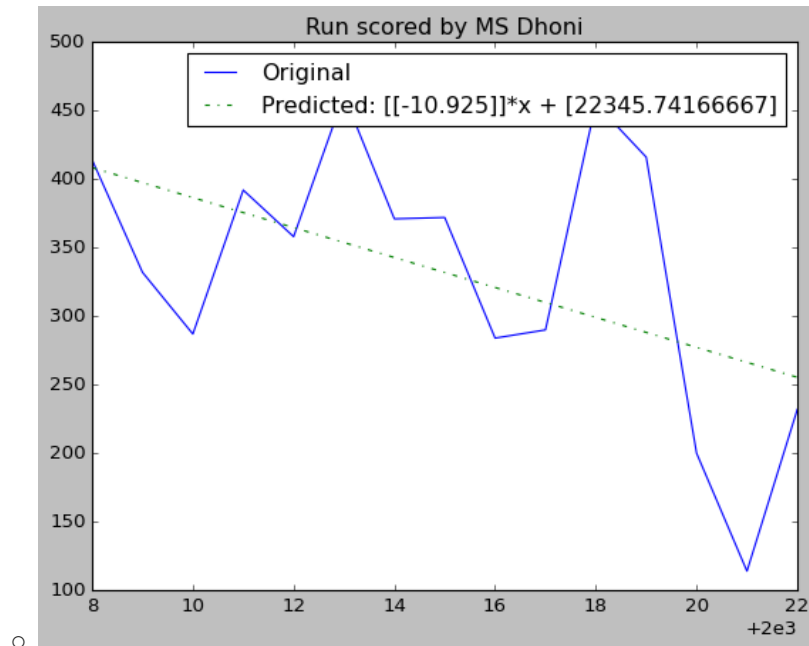


Fig 2.5 : Regression line for MS Dhoni to predict runs

● **For Rohit Sharma:**

- Regression Equation is:  $(-6.4571) \cdot x + (13403.2095)$
- Predicted Outcome for 2023 is: 340.4095 ~ 340 runs.

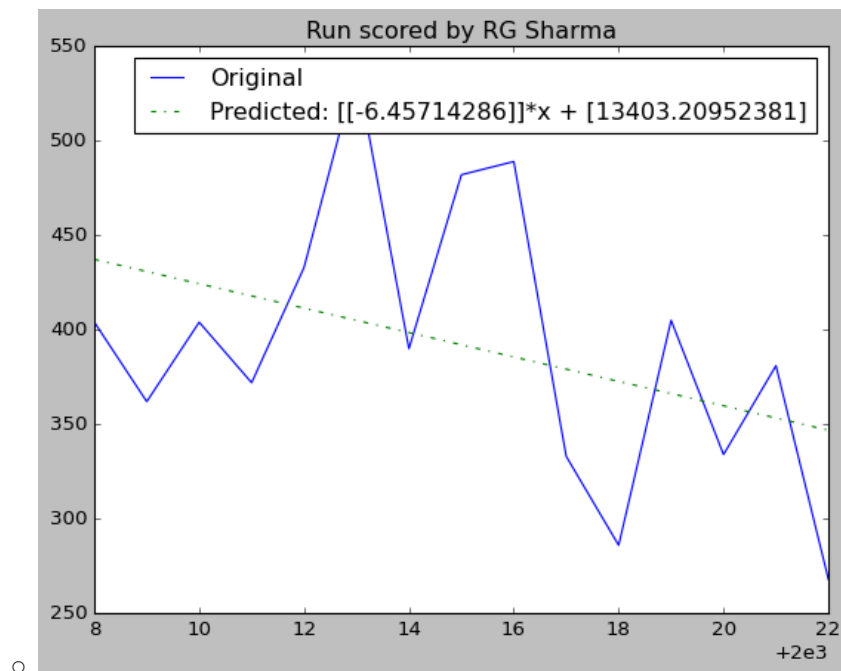


Fig 2.6 : Regression line for Rohit Sharma to predict runs

---

## References

1. Pandas documentation <https://pandas.pydata.org/docs/>
2. Scikit-learn documentation <https://scikit-learn.org/stable/>
3. Hands-On Exploratory Data Analysis with Python, by Suresh Kumar Mukhiya, Usman Ahmed
4. Machine Learning, A Probabilistic Perspective by Kevin P. Murphy
5. Dataset: [https://cricsheet.org/downloads/ipl\\_csv2.zip](https://cricsheet.org/downloads/ipl_csv2.zip)

---

**CV**



Dhirubhai Ambani  
Institute of Information and Communication Technology

# ROHAN SHAH

M.sc., Data Science

## EDUCATION

Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT) (M.sc., Data Science)

CPI: 8

August 2022 - Present    Gandhinagar, Gujarat

Vishwakarma Government Engineering College (GTU) (B.E., Electronics and Communication Engineering)

CPI: 7.86

August 2017 - June 2021    Ahmedabad, Gujarat

Nirman High School (GSEB)

Percentage: 80%

2017    Ahmedabad, Gujarat

Sheth C.N. Vidyalaya, High School (GSEB)

Percentage: 83%

2015    Ahmedabad, Gujarat

## SKILLS

Area(s) of Interest : DBMS, Python, Data Visualization

Programming Languages : Python, R

Tools and Technologies : Jupyter Notebook, Google Colab, Rstudio, Postgres

Technical Electives : Exploratory Data Analysis

## POSITIONS OF RESPONSIBILITY

Event Manager

Managed an event in GTU Central Techfest

April 2019    Ahmedabad, Gujarat

Volunteer

Run for Green Environment

April 2018    Ahmedabad, Gujarat

## INTERESTS

- Sports
- Travelling
- Reading

## PROJECTS

Netflix Data Analysis:

August 2022 - December 2022

- Performed the initial part of the life-cycle of a data science project which included data cleaning, EDA, data manipulation, data filtering, and data visualization.

- Analyzed the trend of Netflix shows and movies, future number of movies and shows, and number of audience views.

- Guide: Nishith Kotak

Waiter Tip Prediction:

March 2022 - June 2022

- Hands-on experience with visualization libraries such as plotly, matplotlib to represent this data graphically, which helped us identify patterns and trends in the data.

- Collected and Cleaned data from various sources and used machine learning algorithms such as regression to build and evaluate predictive models.

Gesture Control Car:

August 2020 - July 2021

- Established serial communication with arduino uno wifi and raspberry Pi3 and used MPU 6050 for 6-axis motion.

- Used Python to develop the code efficiently and Arduino to control the car's hardware components, such as the motors and sensors.

- Guide: Rahul Patel

## ACHIEVEMENTS

- Google Data Analytics Professional Certificate (Coursera)



Dhirubhai Ambani  
Institute of Information and Communication Technology

# KARAN PARASHAR

MSc. Data Science

## EDUCATION

Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT)

CPI: 8.44

August 2022 - Present    Gandhinagar, Gujarat

St. Xavier's College

CPI: 7.67

July 2018 - August 2021    Ahmedabad, Gujarat

Terf English Medium School (GHSEB)

Percentage: 71.54%

2016 - 2018    Ahmedabad, Gujarat

D.P. High School (GSEB)

Percentage: 86.84%

2006 - 2016    Ahmedabad, Gujarat

## SKILLS

Area(s) of Interest : Data Science and Exploratory Data Analysis etc.

Programming Languages : C, Python, R and MATLAB

Tools and Technologies : Jupyter Notebook, Google Colab, MySQL, PostgreSQL

Technical Electives : EDA

## POSITIONS OF RESPONSIBILITY

Student Volunteer

Institute for Plasma Research, Bhat, Gandhinagar.

December 2018 - February 2019

Project Volunteer

On the occasion of National Science Day in 2019, I worked as a project volunteer at the Institute for Plasma Research, to prepare a working model showing the phenomenon of superconductivity.

December 2018 - February 2019

## INTERESTS

- Photography
- Performing Arts

## EXPERIENCE

aMarketForce Pvt. Ltd:

Feb 2022 - Aug 2022

- I worked as a process executive in the company. Majority of my work was around lead generation and B2B marketing for their clients, under the mentorship of assigned TL.
- Guide: Jinel Shah

Harsha Engineers International Limited

Aug 2021 - Dec 2021

- I handled sales and supply chain operations for the distribution of DGBB cages throughout India and to a few foreign clients.
- Guide: Mr. Vishwadeep Zala

## PROJECTS

Netflix Data Analysis:

August 2022 - December 2022

- Performed data cleaning, EDA, data manipulation, data filtering, and data visualisation as the first stage of a data science endeavour.
- Analysed Netflix's programming trends, projected production numbers, and viewership figures.
- Guide: Nishith Kotak

Langmuir Probe Plasma Diagnostics:

May 2019 - March 2020

- Data analysis of plasma parameters in Plasma shots and GDC, by Langmuir Probe method in ADITYA-U Tokamak.
- Guide: Dr. Joydeep Ghosh and Dr. Tanmay Macwan

## ACHIEVEMENTS

- Official Ambassador of International Youth Math Challenge

---

# Mohammed Nauman Shaikh

naumanshaikh2303@gmail.com | (+91) 8401682448

---

## EDUCATION

DHIRUBHAI AMBANI INSTITUTE OF INFORMATION AND COMMUNICATION  
TECHNOLOGY

Gandhinagar, India

*Master of Science in Data Science (MS DS)* CGPA - 8.89/10

St. Xavier's College Ahmedabad, Bachelor of Science specialized in Statistics.

---

## PROJECTS

### Analysis of Import-Export Data of India (2008-2018)

- Performed the initial part of the lifecycle of a data science project which included data cleaning, EDA, data manipulation, data filtering, and data visualization.
- Gained hands-on experience in using Pandas, NumPy, plotly, matplotlib, seaborn, pytrend.
- Analyzed the Trend of Indian import - export and found some meaning full insights.

### Earthquake Prediction using machine learning techniques:

- \* To predict the disastrous natural calamity, first collected data from US government website, also did data cleaning , EDA and data manipulation.
  - \* Applied Machine Learning models and calculated their performance.
- 

## TECHNICAL SKILL SET

**CODING LANGUAGES** - Python (Advanced), R (Intermediate), SQL (Intermediate), C (Beginner)

**SOFTWARES/TOOLS** - Pycharm, Jupyter Notebooks, RStudio, Postgres, SQL, MS Excel

**TECHNICAL KNOWLEDGE** - Data Science, Machine Learning, Data Structures, DBMS

**COMPETITIVE PLATFORM** - HackerRank.

## ADDITIONAL

**HOBBIES** - Football, watching Movies.

## Nisarg Shah

✉ [nisargshah778@gmail.com](mailto:nisargshah778@gmail.com)

☎ : +91 6353 5787 75

### Education Details

Examination /Degree	Year of Passing	Discipline	Aggregate %/ CGPA	Institute	Board/ University
M. Sc.	-	Data Science	9.4 (Sem - I)	Dhirubhai Ambani Institute of Information and Communication Technology	-
B. Sc.	2022	Statistics	8.8	St. Xavier's College	Gujarat University
HSC (12th)	2019	Science (A- group)	65.30%	Sharda Mandir Modern School	G.S.E.B
SSC (10th)	2017	Matric	91.33%	Sharda Mandir Modern School	G.S.E.B

### Professional Skills

- Adept at MS Office (Word, Excel, PowerPoint etc.)
- Notion (Knowledge management/Productivity software)
- Programming Languages : C, Python, R, Java, Windows Batch Scripting
- Data Cleaning, Visualisation, Basic Machine Learning in Python
- Languages : English, Gujarati, Hindi

### Personal Skills

- Dedicated
- Responsible
- Organized
- Teamwork

### Achievements/Publications

- Research Project : Impact of Pandemic On Mental Health of College Students
- Certificate of Merit in Prof. A.R. Rao Mathematics Competition (2019)