Most Runs in Cricket

Date: 29/10/2022

Author: Hurair Mohammad

Introduction

This data is about cricket players with the highest number of scores of a specific player. There are 88 players data in the dataset with different records of the player like highest score, total runs, total sixes etc. From the most run scorer of all time are Indians with most 100’s is well. The data say that Indian and Australian are good cricket players.

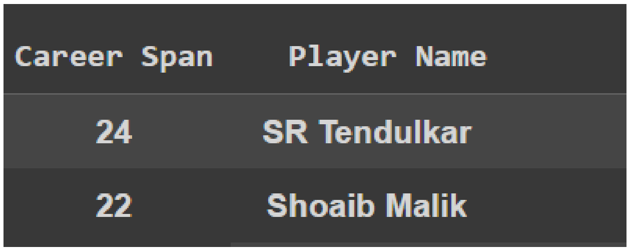
Analysis

Firstly, I check the Data manually. As the heading of columns are easy it is easy what their meaning is and what are them. We have data of different cricket player of the world with good record in the cricket with most runs or most sixes or most average etc. from different countries.

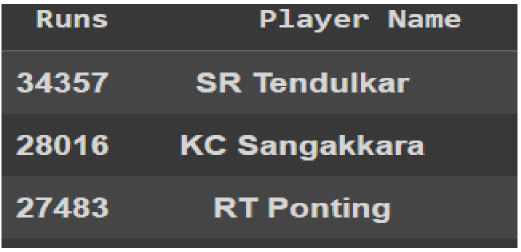
Reading Data from the CSV file I have came to know that the country name and player name are given combine so first I fix it with Introducing a new column of country. In the data there is other column of showing the player starting career year and retirement year I am going to make a new column from this where I subtract retirement year from joining year and it show total year of the player in cricket. After that I checked the data info in if there is any missing value from this, I came to know that we have 88 rows and 20 columns which confirm that the number of players in this dataset is 88. The type of data is in three int64, float64, and object form. The object form data I have to change it into int or float is the requirement.

Knowing that the data has no null values I am going to start the analysis of the data.

1. Finding the Player who play for the longest time and most number of runs score by a player?



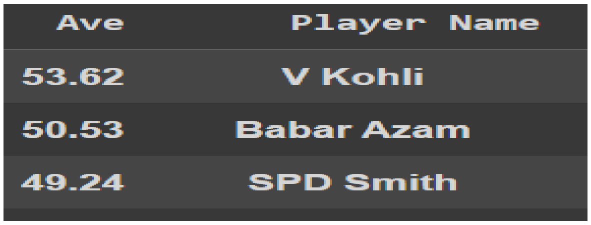
Career Span	Player Name
24	SR Tendulkar
22	Shoaib Malik



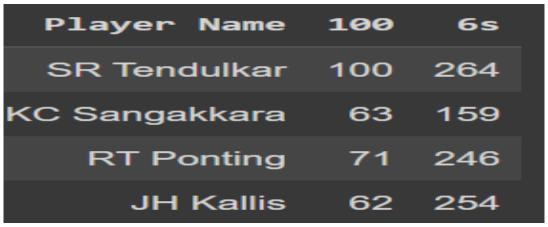
Runs	Player Name
34357	SR Tendulkar
28016	KC Sangakkara
27483	RT Ponting

Tendulkar is the player who play cricket for 24 years followed by Shoaib Malik. In the most run scorer of the list is again Tendulkar who score 34357 runs in his career.

2. The highest average runs scorer in the cricket is most 100s and 6s scorer in cricket.



Ave	Player Name
53.62	V Kohli
50.53	Babar Azam
49.24	SPD Smith



Player Name	100	6s
SR Tendulkar	100	264
KC Sangakkara	63	159
RT Ponting	71	246
JH Kallis	62	254

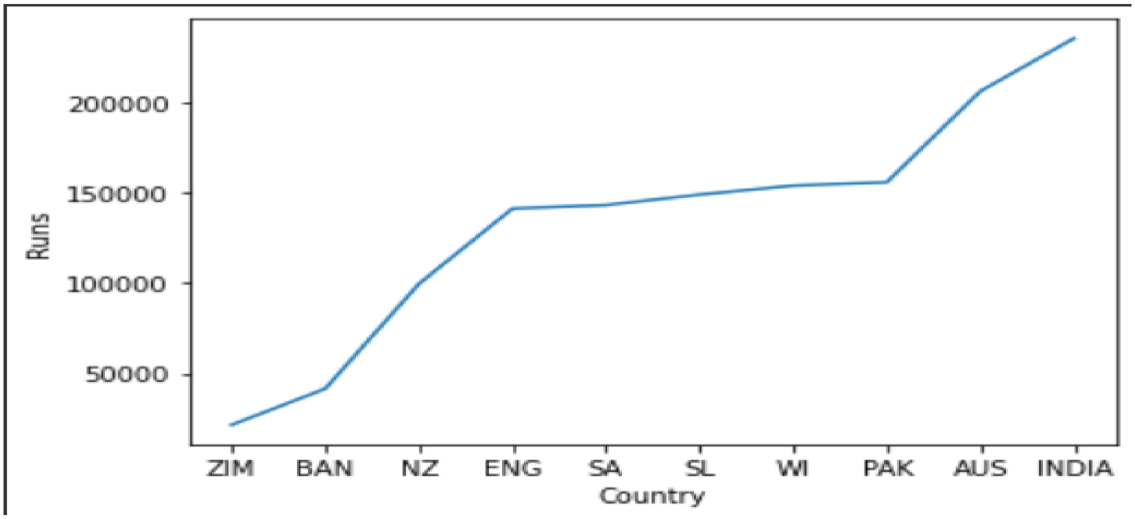
The highest average of the player is coli whose average score is 53.6. and most 100s done.

From the above data we get the most of the record hold by Tendulkar with the most run scorer 24 year of career. In this 24 year of career, he scores 34k plus runs with making the greatest number of 100s and the greatest number of 6s in the list. While highest average of the player is V Kohli who average run score is 53.6. Both of the players are Indian in this we conclude the batting order of Indian batsman is on the top in this dataset.

Visualization

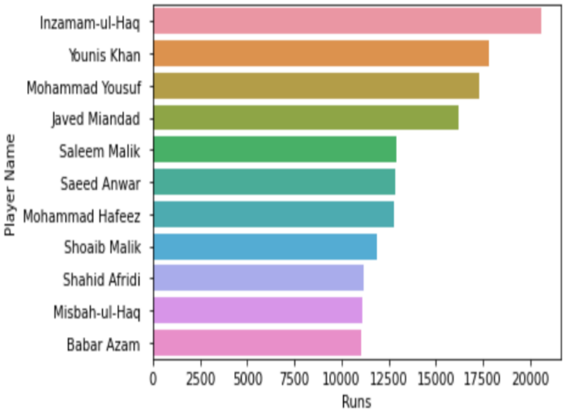
For visualization import visualization packages in python. For visualization I have chosen different point to visualize most of them are visualizing the above data.

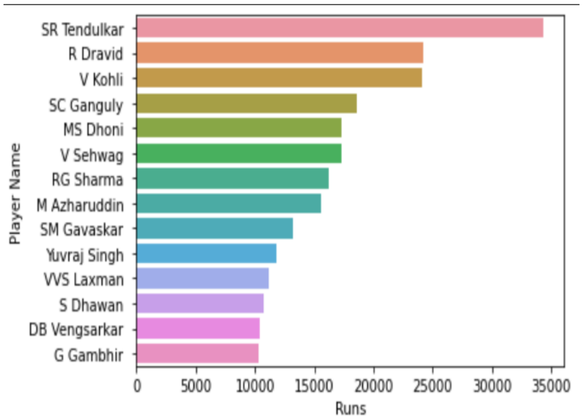
1. Which country players make most numbers of runs?



Is show in the graph that the greatest number of run scores by Indian followed by Australia.

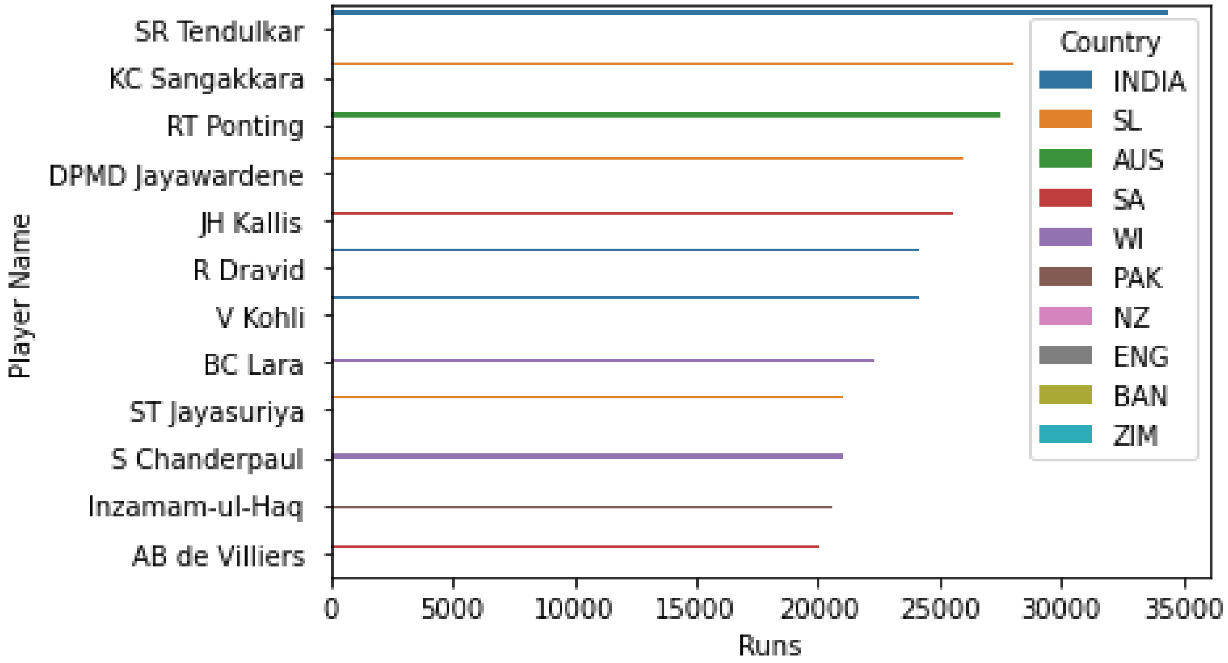
2. Player score runs for Pakistan and India



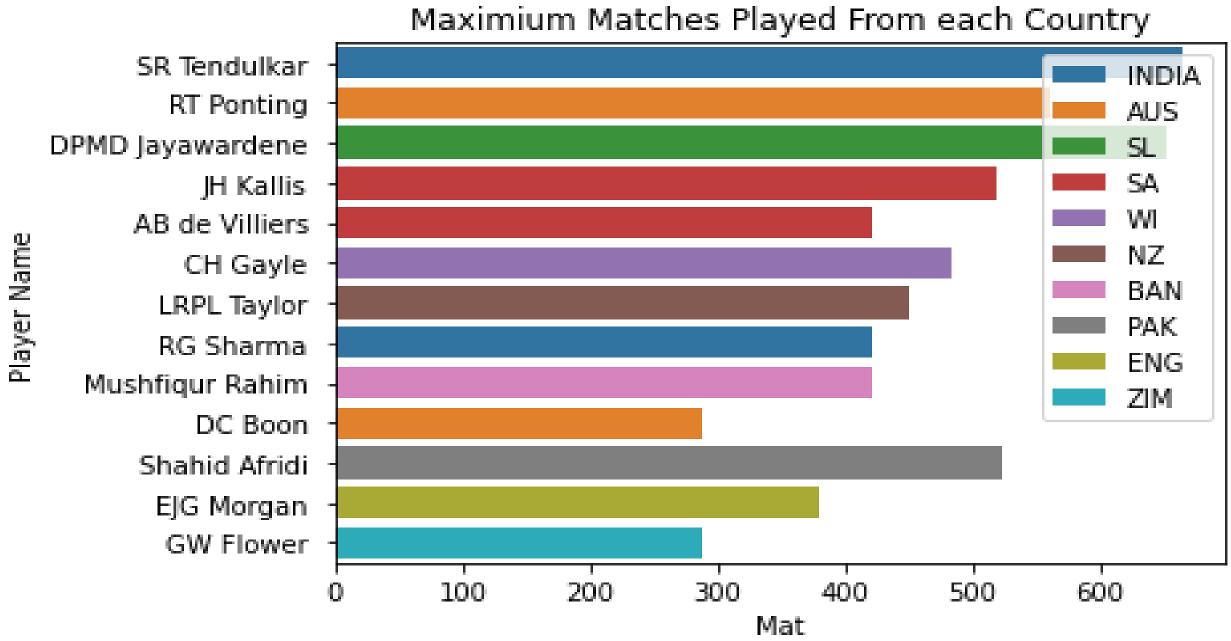


For Pakistan Inzamam-ul-haq is the greatest run scorer for India it is Tendulkar.

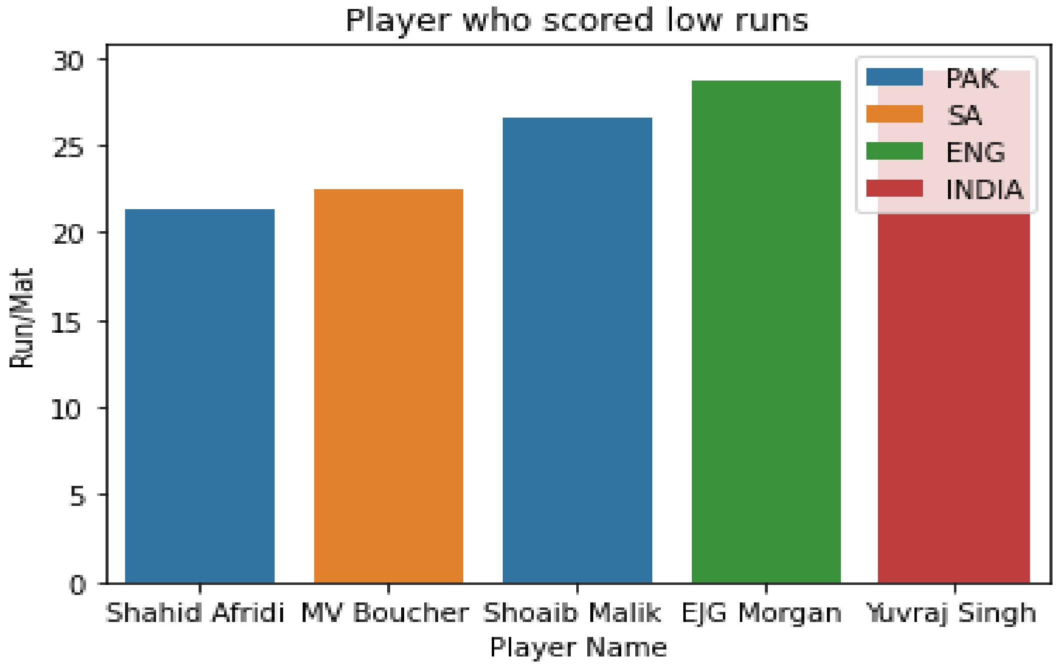
3. Players who score 20,000+ Score



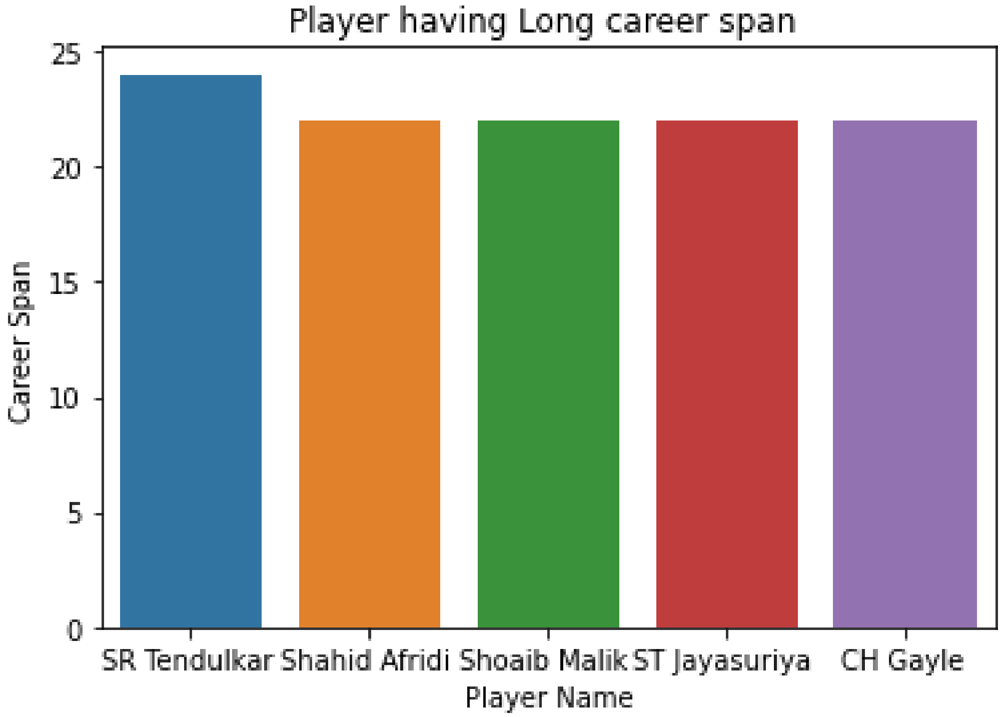
4. Number of Matches Played by a Player



5. Least number of Run score by Players



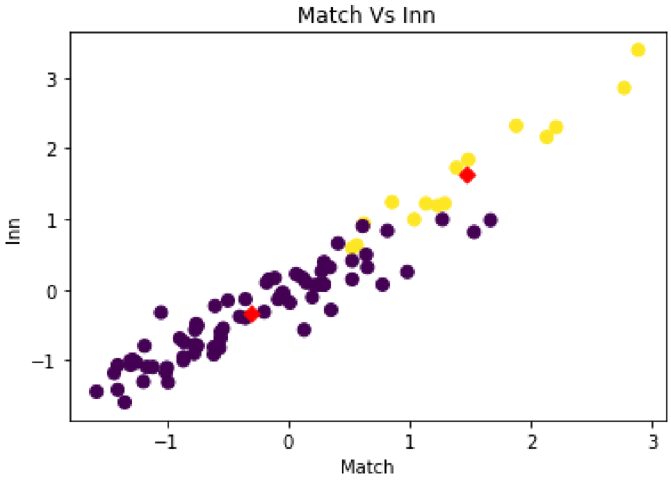
6. Longest Career



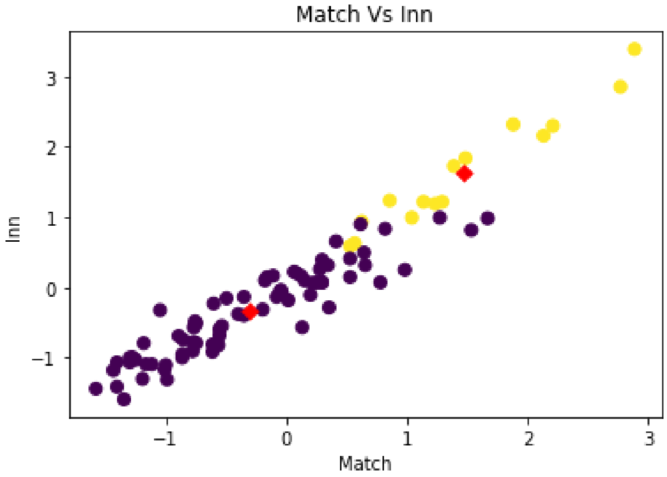
After visualization I have to prepare data for model training. For model training I chose some features which I think will be good. I used K-means algorithm for this data because K-Means Clustering is an unsupervised learning algorithm that is used to solve the clustering problems in machine learning or data science.

First Import K-means from sklearn.cluster then select a cluster range for it where we have to find a good k value on which a least number of cluster error is present. After training I have display different graph for different categories to avoid confusion.

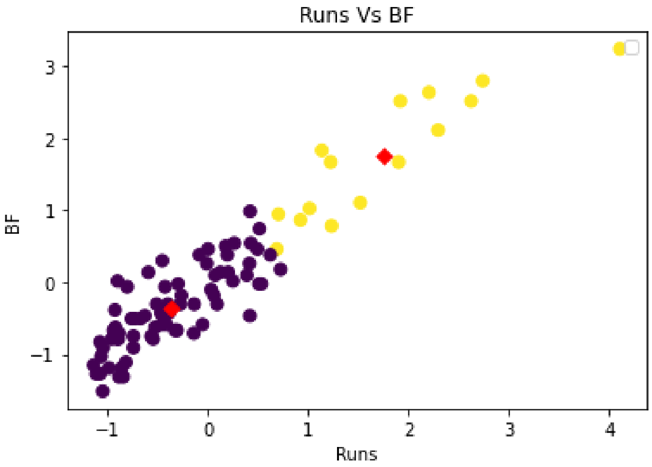
i. Total Match vs Inns



ii. Runs vs Inns

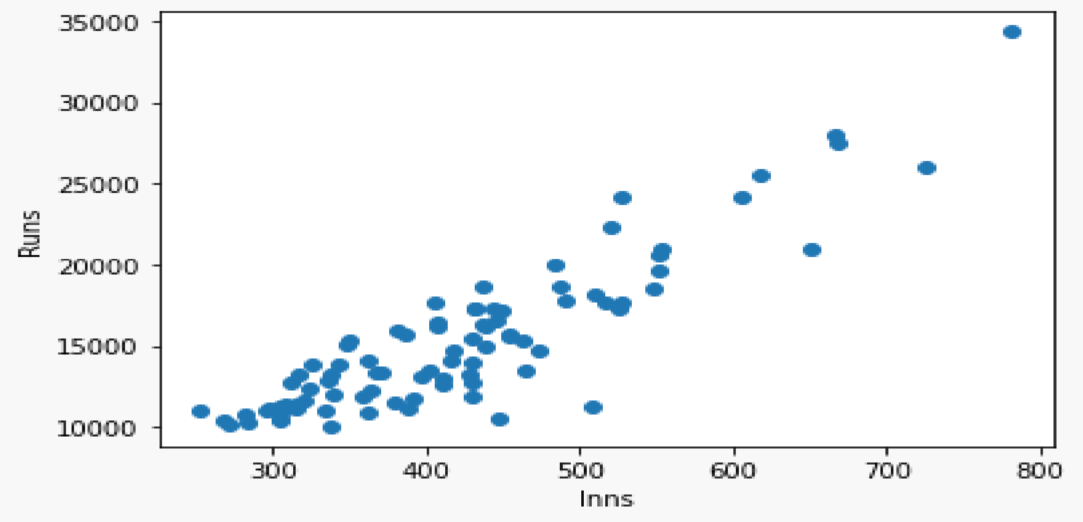


iii. Runs vs BF

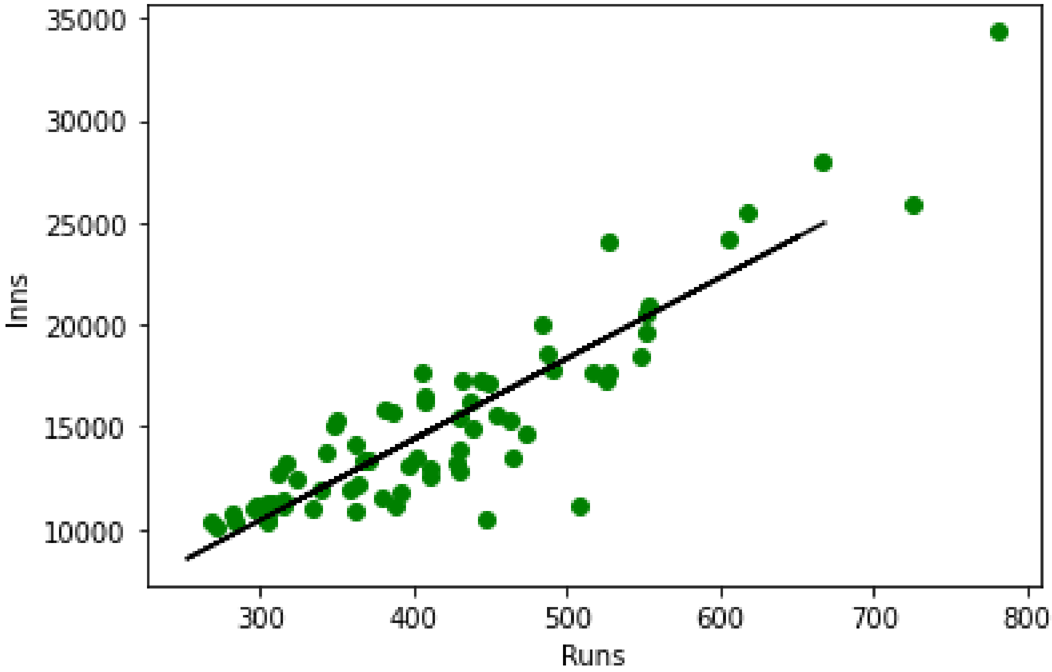


Second model I train is Linear Regression for this model training I chose only one dependent and one independent feature Dependent Feature I chose is Runs and Independent is Inns Where the total runs depend on total innings.

The data visualize



After Training the model then we evaluate it. While evaluating the model with test data we get the accuracy of 78.5% which is good one this data. The graph of the model on data is


Inns
Runs