Cricket Analytics

Abstract

The purpose of this case study is to showcase the factors affecting the runs scored by a batsman during their innings. Most of the time people think that runs scored are solely depended upon the balls faced (quantitative) and form (qualitative). Form is an abstract subject here it is only observed and cannot be expressed in numbers. Here, we are trying to figure what others factors contribute to the runs scored hence get an idea of the form. We have taken Balls Faced, Minutes batted and Strike Rate. The first model includes all of these variables, in the second model we drop one of the variables after running a correlation analysis among the independent variables and consequently run the second model. The best fit model gives us an equation which is used to predict the scores after giving the independent variables. Furthermore, we give individual batting scores against all the opponents and strike rate for a specific batsman which can be used by opposing teams to either formulate a strategy to dismiss the batsman or by the home team to prepare.

Objectives

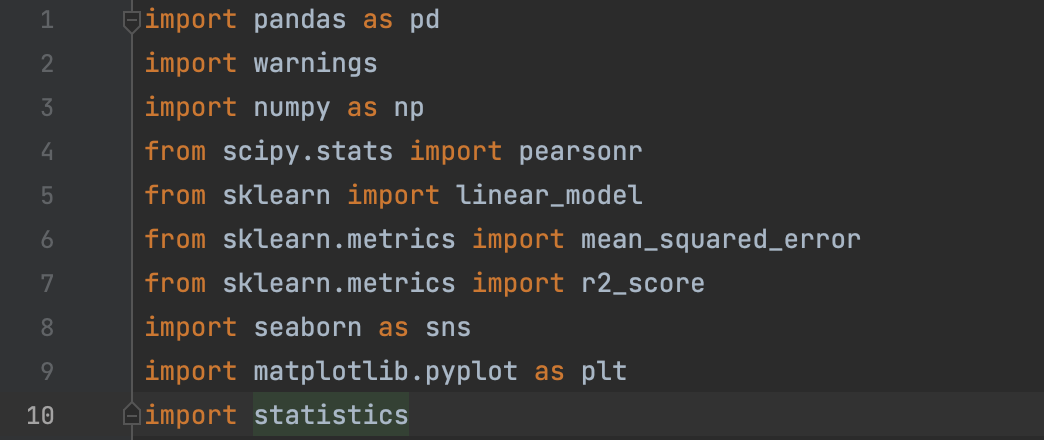
* Factors affecting runs scored by a batsman
* To give stats for a specific batsman based upon the data

Background

The dataset consists of data for players from 1970’s to 2019 and gives their various playing statistics like Runs Scored, Out/Not Out/Balls Faced/Strike Rate/ Minutes spent on the crease/ Innings number/ Opponent Team. The data set consists data from 1970’s to 2019.

|  |  |
| --- | --- |
| **Variable** | **Data Type** |
| Innings Player | object |
| Innings Runs Scored | object |
| Innings Runs Scored Num | int64 |
| Innings Minutes Batted | int64 |
| Innings Batted Flag | float64 |
| Innings Not Out Flag | float64 |
| Innings Balls Faced | int64 |
| Innings Boundary Fours | object |
| Innings Boundary Sixes | object |
| Innings Batting Strike Rate | float64 |
| Innings Number | object |

Packages Used

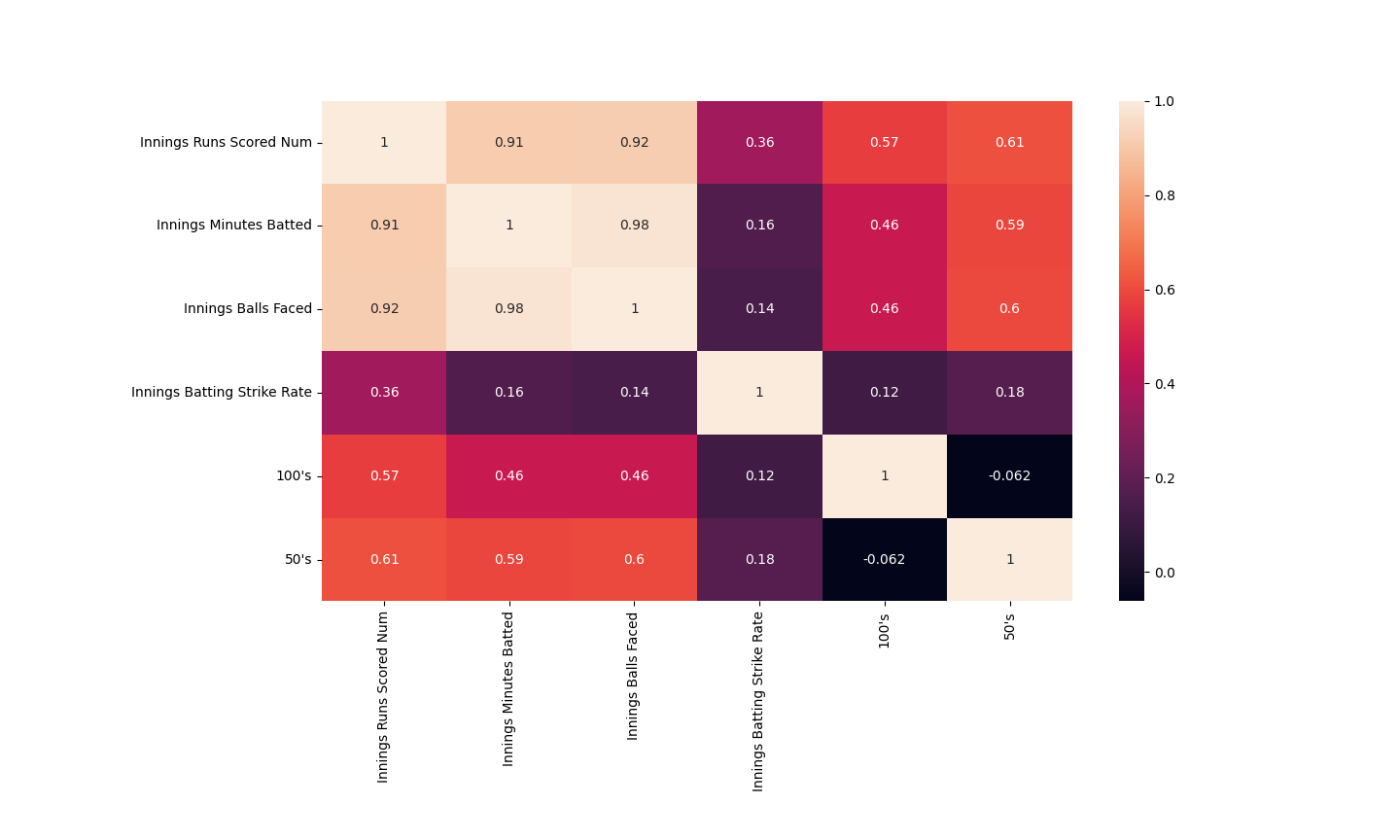


These were the packages used for the analysis

Sk-learn was the package used for linear regression

Pandas and numpy for data handling

**Correlation Chart**



Here we can observe the correlation values between the variables and and based on this we will structure our first linear regression model

**Model-1**

**The model is being prepared with respect to a specific player, the choice is being given to the user to select the player of their choice. For our model we have taken Virat Kohli as the player.**

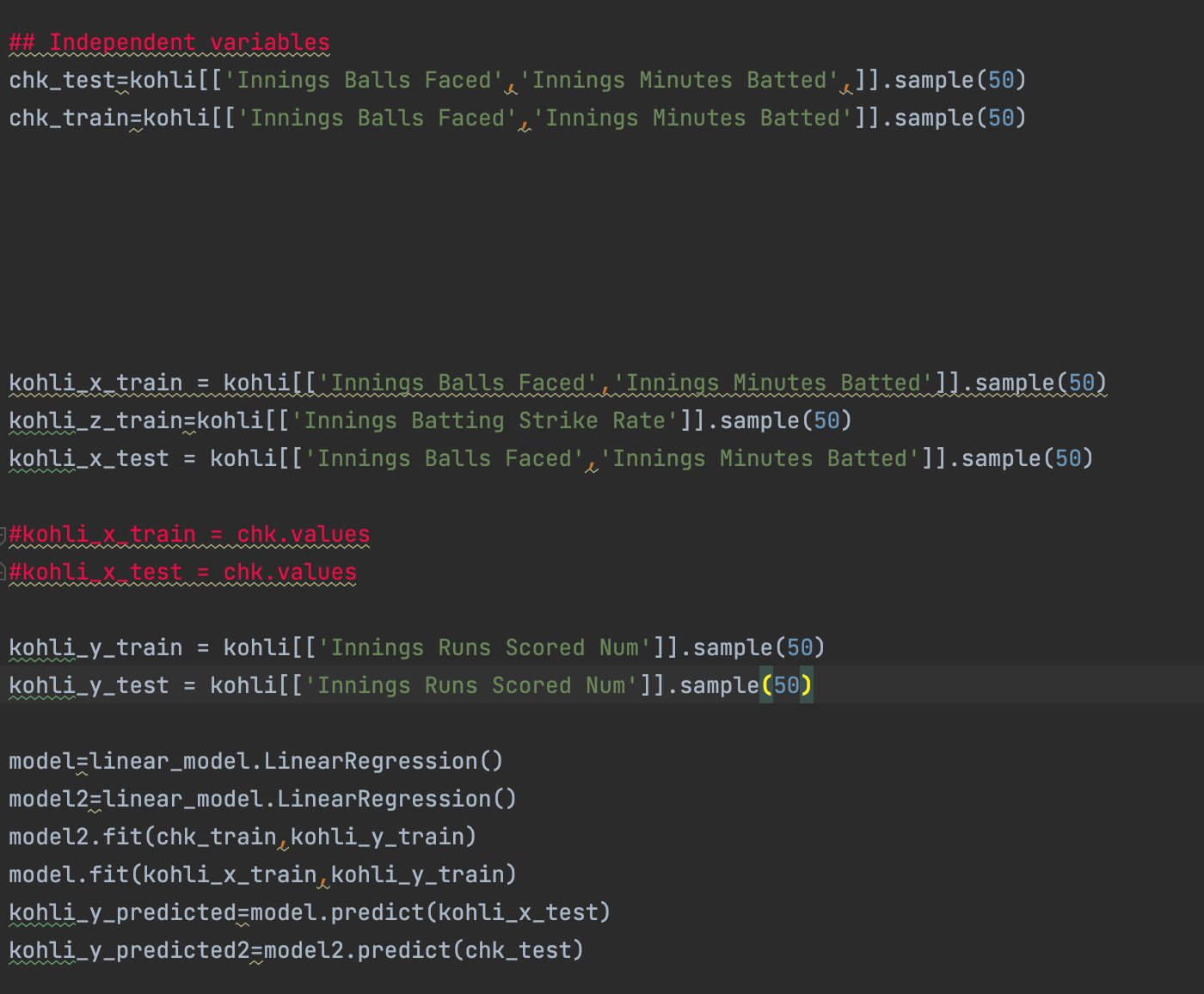
|  |  |
| --- | --- |
|  |  |
| **Model-1** |  |
|  |  |
|  |  |
| **Mean squared error is** | **138.388881** |
| **Weights** | **[[0.72650421, 0.15139869 ,0.21574881]]** |
| **Intercept** | **[-17.83242955]** |
| **R squared** | **0.924535404** |
| **Correlation Coeff** | **0.95980135179828, 9.00899571213922e-110** |
|  |  |
| **Linear Equation** |  |
| **Y=.726(X1)+0.513(X2)+0.215(X3)-17.83** |  |
|  |  |
|  |  |
| **Enter balls faced** |  |
| **35** |  |
| **Enter minutes on crease** |  |
| **20** |  |
| **Enter Strike Rate** |  |
| **95** |  |
| **Predicted score is** | **[[23.86333179]]** |
| **Batting avg.** | **48.30845771** |

**Model 1 takes Balls Faced, Minutes on Crease and Strike Rate as independent variables and based upon that the linear regression model is done, using which we come up with a prediction model that can predict the score. The mean squared error is 138.88**

**And weights are .72,.15,.21 and the intercept are -17.832. The correlation Coeff. Is .95 and p-value is 9.008e-110 which is less than (.05) which can be considered as significant.**

**The equation after which is *( Y=.726(X1)+0.513(X2)+0.215(X3)-17.83).***

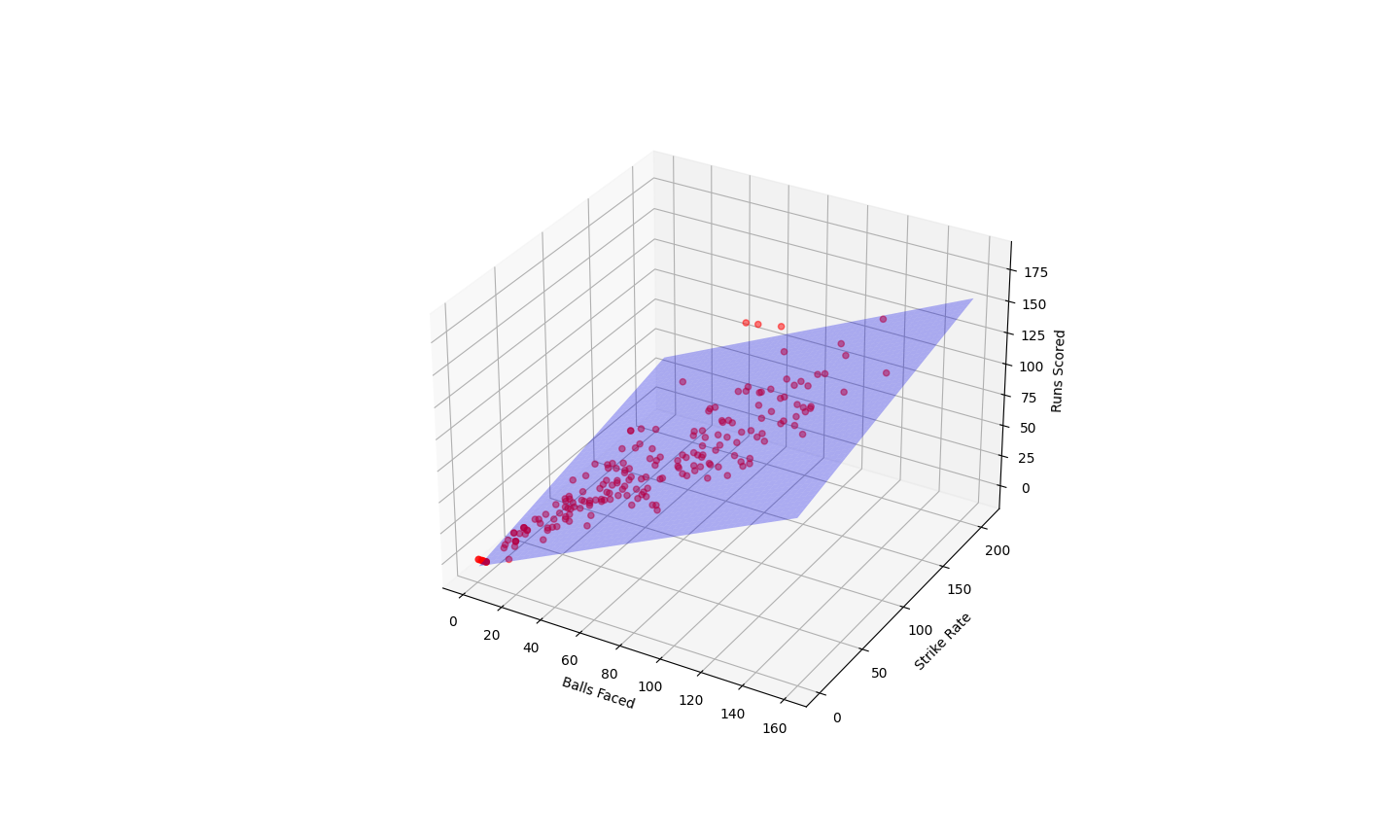
**Code for linear regression**

****

**Here Kohli is the variable used after sub-setting the data-set for a player after which the multivariate linear regression was run.**

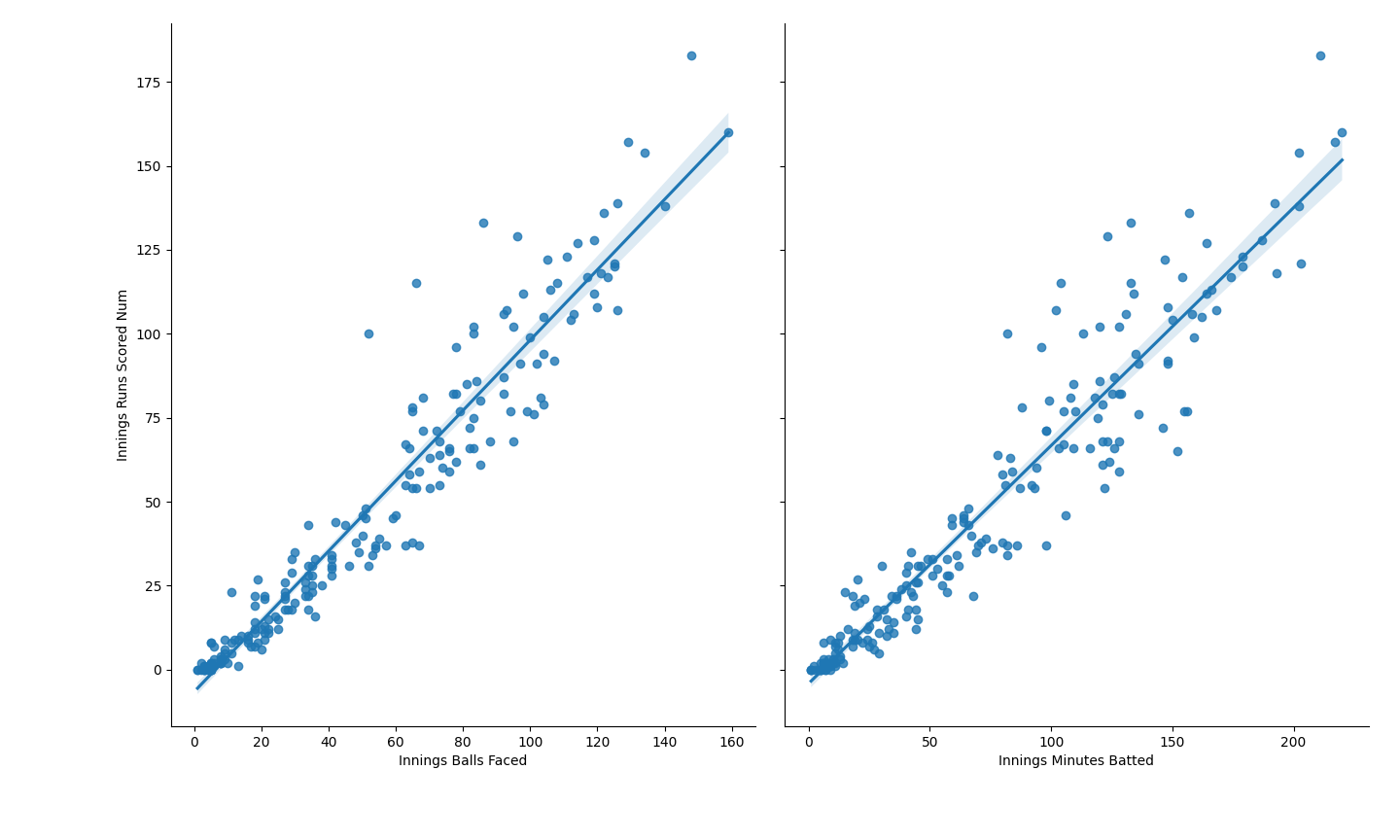
**The data frame was divided into test and train and random values were given to them.**

**3-D Plot**

****

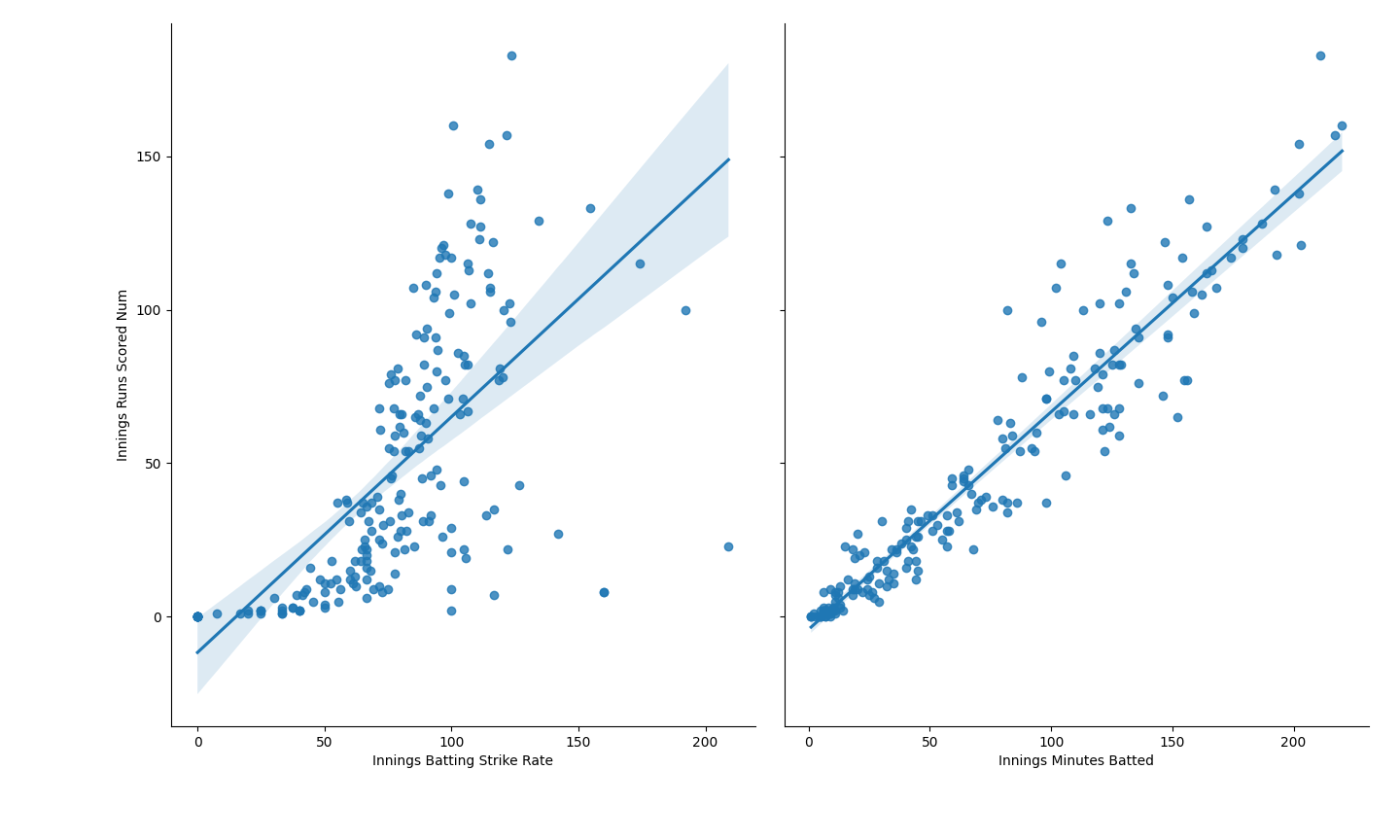
**This 3-D plot takes Balls Faced, Strike Rate, Runs Scored for plotting and the points closer to the plane represents high correlation.**

**Pair-Plot**

****

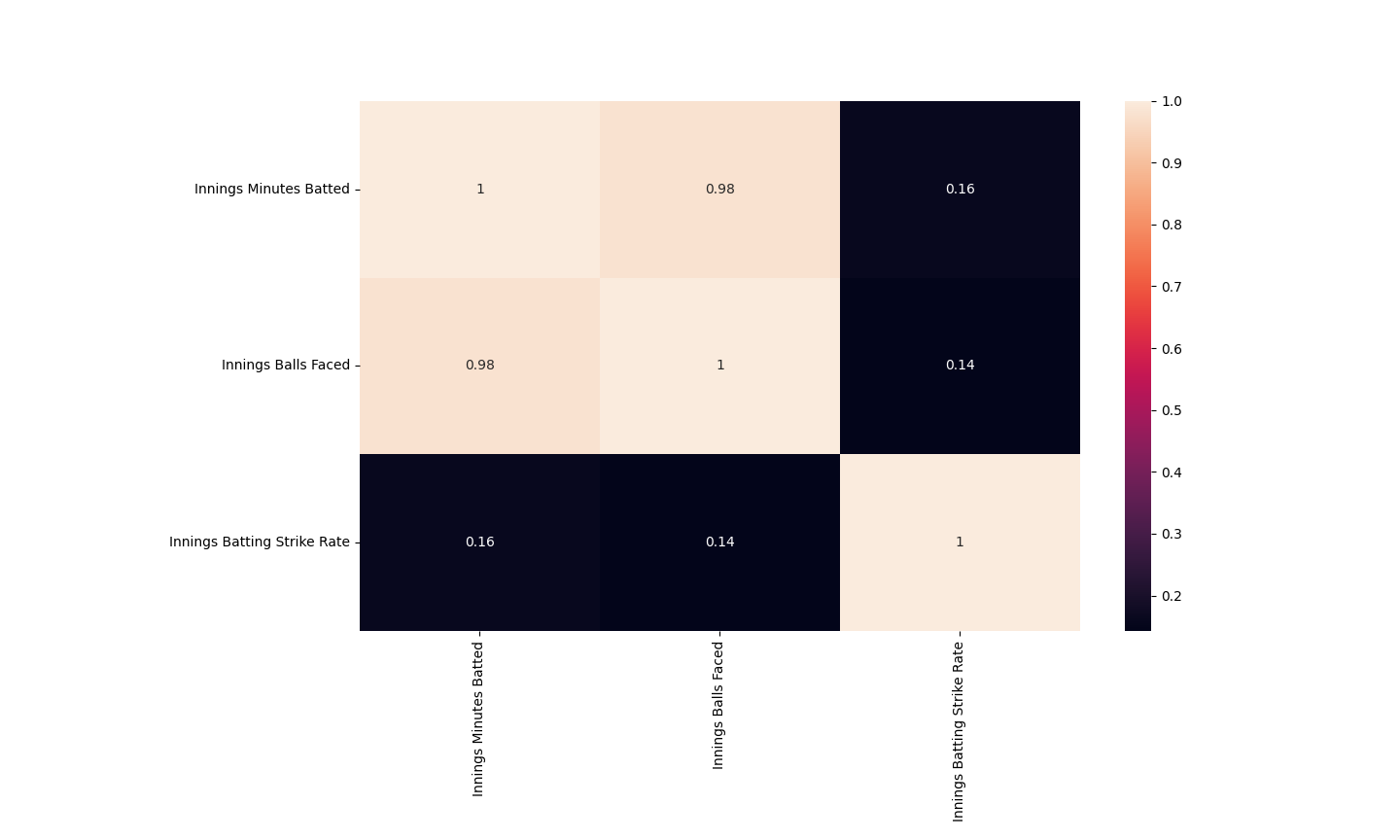
* The plot depicts that Runs Scored are strongly related with Balls Faced and Minutes Batted
* After 60 Balls the deviation tends to increase in terms of balls faced
* After 100 mins the deviation increases in terms of Minutes batted

**Scatter Plot**

****

* The graph shows us that the Innings runs scored are not strongly related with the strike rate of the batsman

**Correlation among independent variable**

****

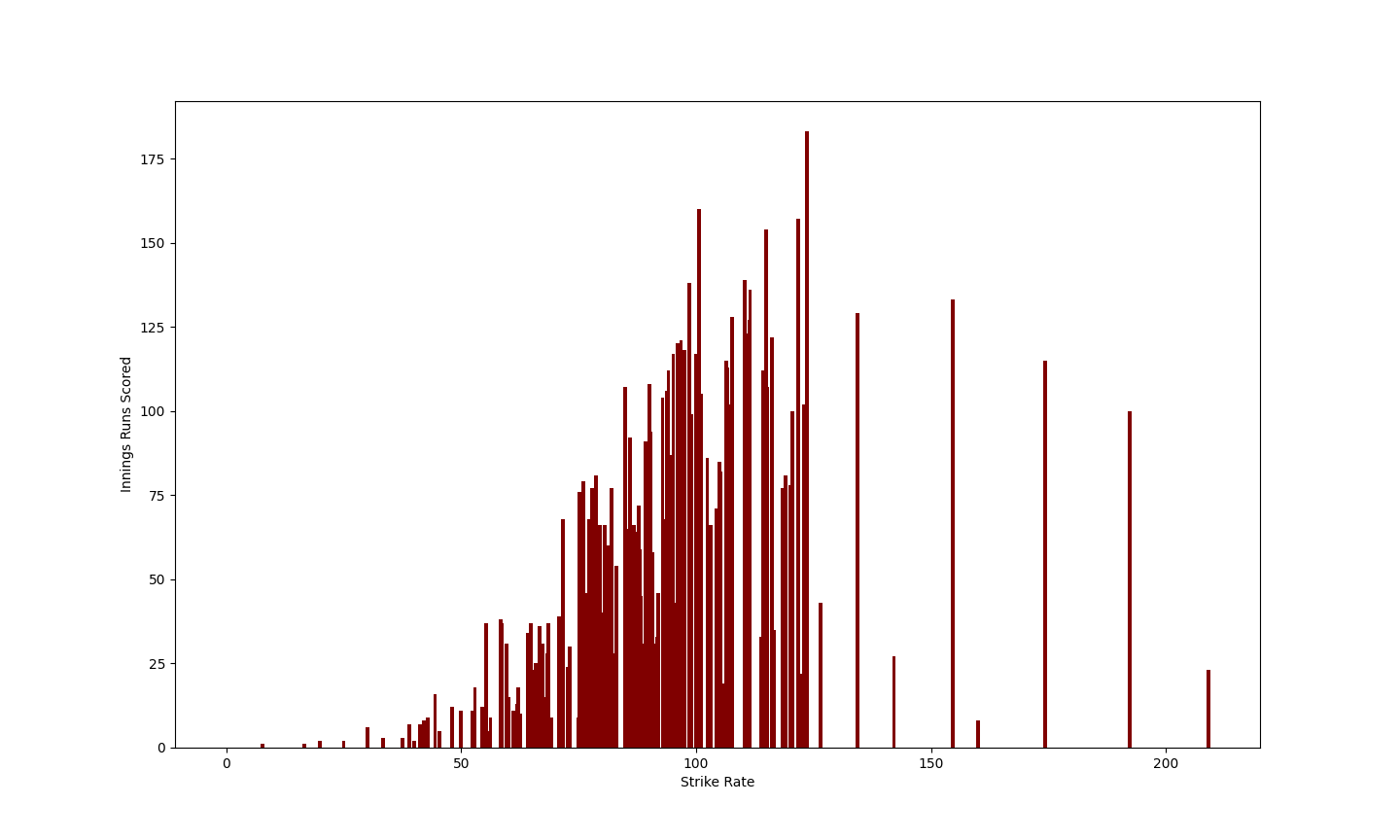
**In this Correlation Analysis we can clearly deduce that Batting Strike Rate has very low correlation with the other variables hence it can be removed in the next model analysis.**

**Model-2**

|  |  |
| --- | --- |
| **Model-2** |  |
|  |  |
| **Mean squared error is** | **96.18649942** |
| **Weights** | **[[0.73957538 0.21404674]]** |
| **Intercept** | **[-6.30274125]** |
| **R squared** | **0.94754871** |
| **Correlation Coeff** | **(0.9598013517982803, 9.19899571213922e-112)** |
|  |  |
|  |  |
| **Enter balls faced** |  |
| **35** |  |
| **Enter minutes on crease** |  |
| **20** |  |
| **Predicted score is** | **[[31.11932828]]** |
| **Batting avg.** | **48.30845771** |
|  |  |
| **Linear Eq:-** |  |
| **Y= .73(X1)+.21(X2)-.6302** |  |

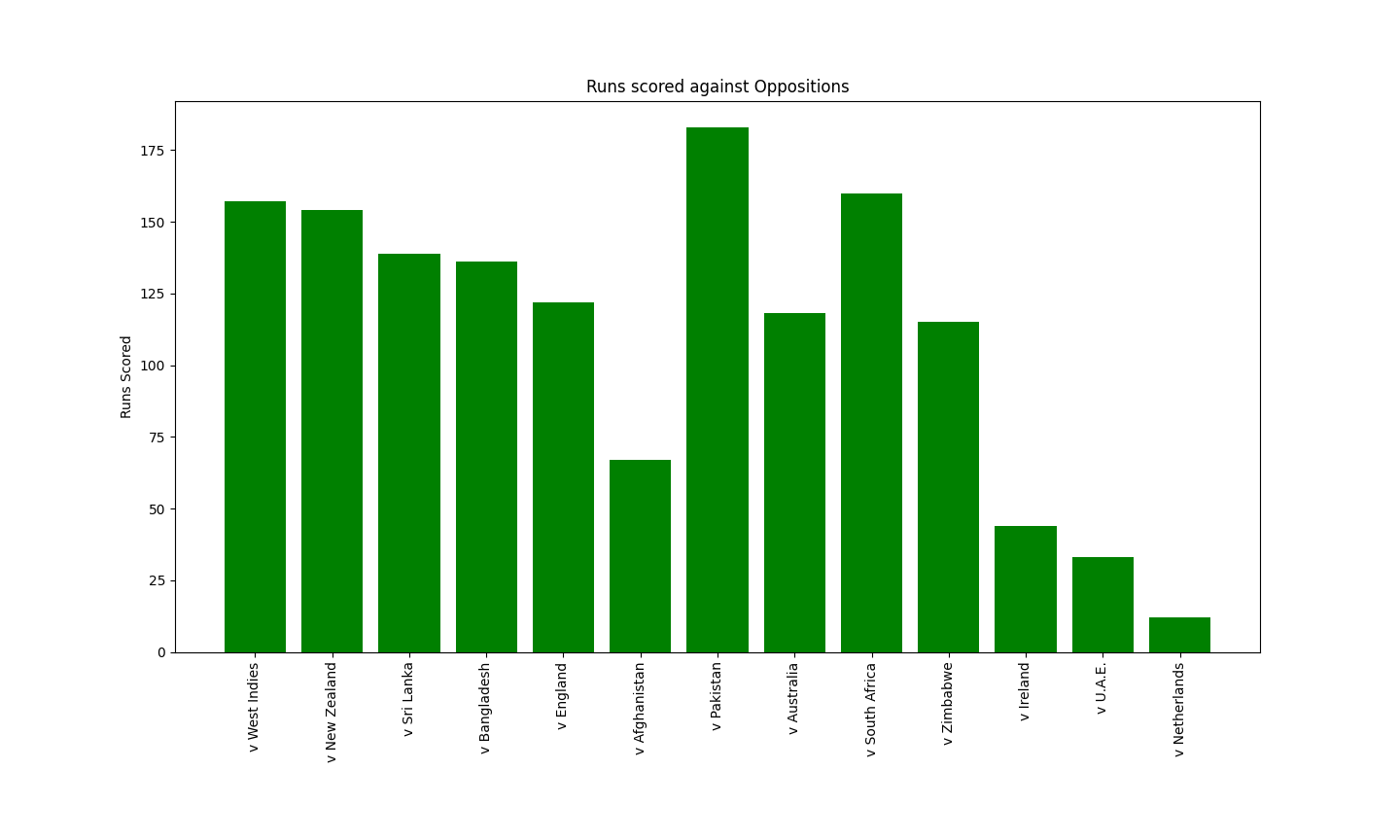
**After running correlation among independent variable, we have eliminated Batting Strike Rate as one of the independent variables. The second model is the best fit model which takes Balls Faced and Minutes Batted as independent variable and Runs Scored as Dependent Variable**

**Bar Graph**

****

**From the graph we can infer that the Majority runs are scored when the strike rate is between 75-115. Thus, when the player has a strike rate in this range, he tends to score the most.**

**Bar Graph**

****

**This graph displays the runs scored by the player against each team. This can be used by opposing teams to better prepare a strategy against the player. Also, the player can also prepare for their opponents.**