Analysis of Cricket Data in R

Tanvi

May 1, 2018

In this post and the next two posts, we will discuss how analytics has revolutionized the world of sports.

First, we will give an overview of the role of analytics in sports and its benefits. Before analytics came into play, the sportsman and the coaches relied on experiences and intuitions to make a decision. The emergence of analytics in sports has not only lead to improving the performance and accuracy of players but also helped in developing better strategies. Today, most of the major teams incorporate analytics. For example, in 2015 the International Cricket Council (ICC) took the help of analytics to analyze the performance of the players and the runs scored by them. They had a partnership with SAP to perform the analysis.

In this post, we will focus on how analytics can be applied to Cricket. The combined data of the bowler as well as the batsman if analyzed efficiently has the ability to give significant insights into the game. This helps the coach as well as the cricketers to make decisions in the right direction. This can be explained in a better way by performing an Exploratory Data Analysis (EDA) on a dataset. The EDA helps in enhancing the experience of the spectators too. This is because the spectators get a detailed view of exactly what is happening in the game.

Analysis in R

To bolster the above points, we will perform an EDA on the datasets of the two most popular players of the Indian Cricket team, Captain Cool -**“MS Dhoni”** and Captain Hot - **“Virat Kohli”**. We will focus on the batting styles of both and then do a comparative analysis of their performance. We will do all of this in R.

We will use the **“cricketr”** package in R. This package scraps the data from the **“ESPN Crickinfo Statsguru”**. As the main focus is on the batting style we will only focus on the variables associated with it. The data variables which are available can be seen below

```
##   Runs Mins BF 4s 6s    SR Pos Dismissal Inns  Opposition        Ground
## 1   12   33 22  1  0 54.54   2       lbw    1 v Sri Lanka      Dambulla
## 2   37   82 67  6  0 55.22   2    caught    2 v Sri Lanka      Dambulla
## 3   25   40 38  4  0 65.78   1   run out    1 v Sri Lanka Colombo (RPS)
## 4   54   87 66  7  0 81.81   1    bowled    1 v Sri Lanka Colombo (RPS)
## 5   31   45 46  3  1 67.39   1       lbw    2 v Sri Lanka Colombo (RPS)
##    Start Date
## 1 18 Aug 2008
## 2 20 Aug 2008
## 3 24 Aug 2008
## 4 27 Aug 2008
## 5 29 Aug 2008
```

The data variables are explained in brief in the order in which they appear above

- the runs faced by the batsman
- the minutes spent at the crease
- the number of balls faced by the batsman
- the number of 4s made by the player
- the number of 6s made by the player

- the strike rate in that particular match
- the position at which the player plays
- the description of dismissal of the player if any
- the innings in which the player played
- the name of the opposition
- the place or the ground on which the match was played
- the date of the match

So with the help of the above variables, we will try to extract as much information as possible. But we will also make sure that the analysis is kept simple so that it is easy for you to understand and implement. We will be dealing with the One Day International(ODI) dataset of Dhoni and Kohli. We will make use of statistics as well as machine learning techniques for performing analysis on the two datasets. We will try to highlight the factors which are responsible for making these two cricketers so popular as well as different from each other.

Importing the Libraries

Let us start by importing the required libraries. Also, we will be explaining you the code in R for Virat Kohli's data mostly but display the outputs for both. With this, you can easily implement the same for Dhoni's dataset.

```
library(cricketr)
library(tidyr)
library(plyr)
library(ggplot2)
library(plotrix)
```

Extraction of Data

Now we will extract the data and store it.We have made use of `getPlayerDataOD` function available in cricketr package to get the data. Here OD signifies One Day. We will be only using the data of One Day International matches of both the players. The parameters used in the function are explained below:

- `profile` signifies the profile number of Virat Kohli which can be found by entering his name on the **"ESPN Crickinfo Statsguru"** website. Profile number of Virat Kohli is *253802* and that of MS Dhoni is *28081*.
- `dir` refers to the directory where you want to store your file
- `file` refers to the file name which you want to specify.
- `type` refers to the type of the data you want.Here type is *batting* as we are interested in getting the batting data only.

Then using the `head` function we have printed the first few rows of Virat's file.

```
virat <- getPlayerDataOD(253802,dir=".", file="kohliOd",type="batting")
dhoni <- getPlayerDataOD(28081,dir=".", file="dhoniOd",type="batting")
```

```
head(virat)
```

```
##   Runs Mins BF 4s 6s     SR Pos Dismissal Inns  Opposition        Ground
## 1   12   33 22  1  0  54.54   2       lbw    1 v Sri Lanka      Dambulla
## 2   37   82 67  6  0  55.22   2    caught    2 v Sri Lanka      Dambulla
## 3   25   40 38  4  0  65.78   1   run out    1 v Sri Lanka Colombo (RPS)
## 4   54   87 66  7  0  81.81   1    bowled    1 v Sri Lanka Colombo (RPS)
## 5   31   45 46  3  1  67.39   1       lbw    2 v Sri Lanka Colombo (RPS)
## 6   2*    6  2  0  0 100.00   7   not out    1 v Sri Lanka Colombo (RPS)
##    Start Date
## 1 18 Aug 2008
## 2 20 Aug 2008
## 3 24 Aug 2008
## 4 27 Aug 2008
## 5 29 Aug 2008
## 6 14 Sep 2009
```

As you can see above there are 12 data variables available in the dataset and all have been explained above.

Cleaning the data

Before performing EDA we will first clean the data and remove all the rows which are not required. For example, in the 6th row above in the Runs Column value is 2*. As we are only interested in the runs scored we need to remove the star from there.

```
virat$Runs <- as.numeric(gsub("\\*","",virat$Runs))
virat$Runs[is.na(virat$Runs)] <- 0
virat<- virat[complete.cases(virat),]
```

This is given to remove the * and only the runs associated has been kept. `*` is used for indicating that batsman was not out. Also after converting it to numeric, all the *not available* entries have been replaced with zero to avoid errors in calculations.

```
head(virat)
```

```
##   Runs Mins BF 4s 6s     SR Pos Dismissal Inns  Opposition        Ground
## 1   12   33 22  1  0  54.54   2       lbw    1 v Sri Lanka      Dambulla
## 2   37   82 67  6  0  55.22   2    caught    2 v Sri Lanka      Dambulla
## 3   25   40 38  4  0  65.78   1   run out    1 v Sri Lanka Colombo (RPS)
## 4   54   87 66  7  0  81.81   1    bowled    1 v Sri Lanka Colombo (RPS)
## 5   31   45 46  3  1  67.39   1       lbw    2 v Sri Lanka Colombo (RPS)
## 6    2    6  2  0  0 100.00   7   not out    1 v Sri Lanka Colombo (RPS)
##    Start Date
## 1 18 Aug 2008
## 2 20 Aug 2008
## 3 24 Aug 2008
## 4 27 Aug 2008
## 5 29 Aug 2008
## 6 14 Sep 2009
```

As we can see the 6th-row data under the runs column has been cleaned. Similarly, we will remove the rows containing **DNB** which means the player did not bat.

```
virat<- virat[virat$Runs != "DNB",]
```

Similarly, we will remove the rows containing **TDNB** which means the team did not bat.

```
virat <- virat[virat$Runs != "TDNB",]
```

Similarly, for performing the above steps for Dhoni's dataset, we need to perform the following steps:

```
dhoni<- dhoni[dhoni$Runs != "DNB",]
dhoni <- dhoni[dhoni$Runs != "TDNB",]
dhoni$Runs <- as.numeric(gsub("\\*","",dhoni$Runs))
dhoni$Runs[is.na(dhoni$Runs)] <- 0
dhoni<- dhoni[complete.cases(dhoni),]
head(dhoni)
```

```
##   Runs Mins  BF 4s 6s     SR Pos Dismissal Inns   Opposition        Ground
## 1    0    1   1  0  0   0.00   7   run out    1 v Bangladesh    Chittagong
## 2   12   16  11  2  0 109.09   7    caught    2 v Bangladesh         Dhaka
## 3    7    2   2  0  1 350.00   7   not out    1 v Bangladesh         Dhaka
## 4    3    8   7  0  0  42.85   7    caught    1   v Pakistan         Kochi
## 5  148  155 123 15  4 120.32   3    caught    1   v Pakistan Visakhapatnam
## 6   28   36  24  5  0 116.66   3    caught    2   v Pakistan    Jamshedpur
##    Start Date
## 1 23 Dec 2004
## 2 26 Dec 2004
## 3 27 Dec 2004
## 4  2 Apr 2005
## 5  5 Apr 2005
## 6  9 Apr 2005
```

You can see from the output that the data has been cleaned.

Now, the data is ready for EDA as all the extraction and cleaning of the data has been done properly. In the next post, we will analyze the performances of the two cricketers in R with this transformed data and do a comparative analysis with the help of graphical representations. Please feel free to comment or ask for more details in the comment section.

Thanks,

Tanvi (https://www.linkedin.com/in/tanvipareek)