Beaten by sheer pace!

Cricket analytics with yorkr!

Tinniam V Ganesh

CONTENTS

Preface

'yorkr' happens to be my second R package for analyzing performances of cricketers. My first cricket package was 'cricketr', which had a really good reception.

The idea of creating this R package, based on data from Cricsheet (http://www.cricsheet.org), came to me, many months back while I was creating my first R package cricketr. I wasn't too sure about how the package would turn out. The challenge with the data in Cricsheet, was the fact that it was in 'YAML' format.YAML is a recursive acronym for "YAML A'int Markup Language". The conversion of yaml to dataframes, was the biggest hurdle in the creation of the package, as I had to maintain the ball-by-ball details, details of extras, non-striker details and how the player got out. In any case, it was great fun and a process of discovery for me, as I let R discover great batting partnerships,, bowling performances against an opposition or all oppositions, and analyze individual batting and bowling performances etc.

The creation of yorkr was a great experience for me. During the course of the package creation I felt like a writer, a painter, a sculptor and a marathon runner. Here's why

- A writer, because when the idea (plot) of the R package based on (cricsheet.org) was formed, I had to create the associated functions (characters of the plot) and their role in the story.
- A painter,who had to change the shades of the functions to make the output plots/data more informative, compact and useful to the end user
- A sculptor, as I chipped and chiseled the functions to its final shape
- But mostly like marathon runner, of whom I have seen a few up close. Once I started to code the package, and created a few functions, I realized that I had reached a point of no-return, and the only way… was forward. Like a marathon runner, I had to put one step in front of another, and keep driving myself to complete the ideas which came to me in a torrent.

The yorkr package, now has 70 functions. Most of the heavy lifting is done by R packages dplyr, ggplot2.I have been amazed at the interesting results that I got out with my yorkr functions. I am certain may interesting insights abound in the data created by the yorkr package, and available in yorkrData.

With this book, I think I have corked the yorkr genie, and flung it into the world wide sea of the internet. Hopefully, when some intrepid reader uncorks the bottle, the released yorkr genie will grant the reader his or her proverbial 3 boons!

Feel free to contact me with your comments, suggestions and thoughts.

Tinniam V Ganesh

6 May 2016

tvganesh.85@gmail.com

1.Introducing cricket package yorkr: Part 1- Beaten by sheer pace!

"We need to regard statistical intuition with proper suspicion and replace impression formation by computation wherever possible"
"We are pattern seekers, believers in a coherent world"
"The hot hand is entirely in the eyes of the beholders, who are consistently" "too quick to perceive order and causality in randomness. The hot hand is a" "massive and widespread cognitive illusion"

```
"Thinking, Fast and Slow - Daniel Kahneman"
```

Introduction

Yorker (noun):A yorker is a bowling delivery in cricket that pitches at or around the batsman's toes. Also known as 'toe crusher'

- My package 'yorkr' is now available on CRAN. This package is based on data from Cricsheet (http://cricsheet.org/). Cricsheet has the data of ODIs, Test, Twenty20 and IPL matches as yaml files. The yorkr package provides functions to convert the yaml files to more easily R consumable entities, namely dataframes. In fact all ODI matches have already been converted and are available for use at yorkrData (https://github.com/tvganesh/yorkrData). However as future matches are added to Cricsheet, you will have to convert the match files yourself. More details below.

You can clone/fork the code for the package yorkr from Github at at yorkr (https://github.com/tvganesh/yorkr)

1. First things first

1. yorkr currently has a total 70 functions as of now. I have intentionally avoided abbreviating function names by dropping vowels, as is the usual practice in coding, because the resulting abbreviated names created would be very difficult to remember, and use. So instead of naming a function as *tmBmenPrtshpOppnAllMtches()*, I have used the longer form for e.g.*teamBatsmenPartnershipOppnAllmatches()*, which is much clearer. The longer form will be more intuitive. Moreover RStudio prompts for the different functions which have the same prefix and one does not need to type in the entire function name.
2. The package yorkr has 4 classes of functions

- Class 1- Team performances in a match
- Class 2- Team performances in all matches against a single opposition (e.g. all matches of India vs. Australia or all matches of England vs. Pakistan etc.)
- Class 3- Team performance in all matches against all Opposition (India vs. All, Pakistan vs. All etc.)
- Class 4- Individual performances of batsmen and bowlers

In this post I will be looking into Class 1 functions, namely the performances of opposing teams in a single match

The list of functions is

1. teamBattingScorecardMatch()
2. teamBatsmenPartnershipMatch()
3. teamBatsmenVsBowlersMatch()
4. teamBowlingScorecardMatch()
5. teamBowlingWicketKindMatch()
6. teamBowlingWicketRunsMatch()
7. teamBowlingWicketRunsMatch()
8. teamBowlingWicketMatch()

9. teamBowlersVsBatsmenMatch()
10. matchWormGraph()

2. Install the package from CRAN

```
library(yorkr)

rm(list=ls())
```

3. Convert and save yaml file to dataframe

This function will convert a yaml file in the format as specified in Cricsheet to dataframe. This will be saved as as RData file in the target directory. The name of the file will have the following format team1-team2-date.RData. This is seen below.

```
convertYaml2RDataframe("225171.yaml","./source","./data")
## [1] "./source/225171.yaml"
## [1] "first loop"

## [1] "second loop"
setwd("./data")

dir()
## [1] "Australia-India-2012-02-12.RData"
## [2] "Bangladesh-Zimbabwe-2009-10-27.RData"
## [3] "convertedFiles.txt"
## [4] "England-New Zealand-2007-01-30.RData"
## [5] "Ireland-England-2006-06-13.RData"
## [6] "Pakistan-South Africa-2013-11-08.RData"

## [7] "Sri Lanka-West Indies-2011-02-06.RData"

setwd("..")
```

4. Convert and save all yaml files to dataframes

This function will convert all yaml files from a source directory to dataframes and save it in the target directory with the names as mentioned above.

```
convertAllYaml2RDataframes("./source","./data")
## [1] 1
## i= 1   file= ./source/225171.yaml
## [1] "first loop"
## [1] "second loop"

## [1] 633  25
```

5. yorkrData – A Github repository

Cricsheet has ODI matches from 2006. There are a total of 1167 ODI matches(files) out of which 34 yaml files had format problems and were skipped. Incidentally I have already converted the 1133 yaml files in the ODI directory of Cricsheet to dataframes and saved then as RData. The rest of the yaml files have already been converted to RData and are available for use. All the converted RData files can be accessed from my Github link yorkrData under the folder ODI-matches. You will need to use the functions to convert new match files, as they are added to Cricsheet. There is also a file named 'convertedFiles' which will have the name of the original file and the converted file as below

convertedFiles

- 225171.yaml:Ireland-England-2006-06-13.RData
- 225245.yaml:England-Pakistan-2006-08-30.RData
- 225246.yaml:England-Pakistan-2006-09-02.RData …

You can download the zip of the files and use it directly in the functions as follows

Note 1: *The package in its current form handles ODIs, T20s and IPL T20 matches*
Note 2: *The link to the converted data frames have been provided above. The dataframes are around 600 rows x 25 columns. In this post I have created 10 functions that analyze team performances in a match. However you are free to slice and dice the dataframe in any way you like. If you do come up with interesting analyses, please do attribute the source of the data to Cricsheet, and my package yorkr and my blog. I would appreciate it if you could send me a note.*
.

6. Load the match data as dataframes

As mentioned above in this post I will use the functions from Class 1. For this post I will be using the match data from 5 random matches between 10 different opposing teams/countries. For this I will directly use the converted RData files rather than getting the data through the getMatchDetails()

With the RData we can load the data in 2 ways

A. With getMatchDetails()

1. With getMatchDetails() using the 2 teams and the date on which the match occurred

```
aus_ind  <-  getMatchDetails("Australia","India","2012-02-12",dir="./data")
```

or

B.Directly load RData into your code.

The match details will be loaded into a dataframe called **'overs**' which you can assign to a suitable name as below
The randomly selected matches are

- Australia vs. India – 2012-02-12, Adelaide
- England vs. New Zealand – 2007-01-30, Perth
- Pakistan vs. South Africa – 2013-07-08, UAE
- Sri Lanka vs. West Indies -2011-02-06, Colombo(SSC)
- Bangladesh vs. Zimbabwe -2009-10-27, Dhaka

Directly load RData from file

```
load("./data/Australia-India-2012-02-12.RData")
aus_ind  <-  overs
load("./data/England-New Zealand-2007-01-30.RData")
eng_nz  <-  overs
load("./data/Pakistan-South Africa-2013-11-08.RData")
pak_sa  <-  overs
load("./data/Sri Lanka-West Indies-2011-02-06.RData")
sl_wi<-  overs
load("./data/Bangladesh-Zimbabwe-2009-10-27.RData")

ban_zim  <-  overs
```

7. Team batting scorecard

Compute and display the batting scorecard of the teams in the match. The top batsmen in are G

Gambhir(Ind), PJ Forrest(Aus), Q De Kock(SA) and KC Sangakkara(SL)

teamBattingScorecardMatch(aus_ind,'India')

```
## Total= 258
## Source: local data frame [8 x 5]
##
##     batsman balls Played fours sixes  runs
##       (fctr)       (int) (dbl) (dbl) (dbl)
## 1 G Gambhir         110     7     0    92
## 2  V Sehwag          20     3     0    20
## 3   V Kohli          28     1     0    18
## 4 RG Sharma          41     1     1    33
## 5  SK Raina          30     3     1    38
## 6  MS Dhoni          57     0     1    44
## 7 RA Jadeja           8     0     0    12

## 8  R Ashwin           2     0     0     1
```

teamBattingScorecardMatch(aus_ind,'Australia')

```
## Total= 260
## Source: local data frame [9 x 5]
##
##        batsman balls Played fours sixes  runs
##          (fctr)       (int) (dbl) (dbl) (dbl)
## 1    DA Warner          23     2     0    18
## 2   RT Ponting          13     1     0     6
## 3    MJ Clarke          43     5     0    38
## 4   PJ Forrest          83     5     2    66
## 5    DJ Hussey          76     5     0    72
## 6 DT Christian          36     2     0    39
## 7      MS Wade          17     1     0    16
## 8    RJ Harris           2     0     0     2

## 9     CJ McKay           3     0     0     3
```

teamBattingScorecardMatch(pak_sa,'South Africa')

```
## Total= 256
## Source: local data frame [7 x 5]
##
##          batsman balls Played fours sixes  runs
##            (fctr)       (int) (dbl) (dbl) (dbl)
## 1      Q de Kock         132     9     1   112
## 2        HM Amla          50     6     0    46
## 3   F du Plessis          21     1     0    10
## 4 AB de Villiers          40     2     0    30
## 5      DA Miller           9     0     0     5
## 6      JP Duminy          20     1     1    25

## 7      R McLaren          21     3     1    28
```

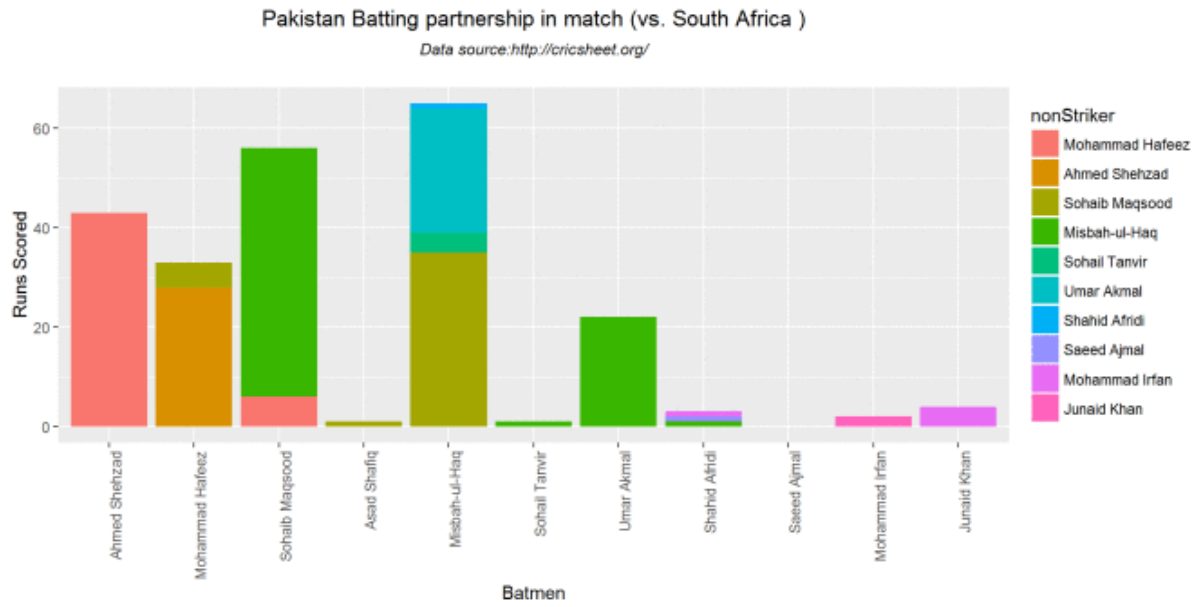
```
teamBattingScorecardMatch(sl_wi,'Sri Lanka')
```

```
## Total= 261
## Source: local data frame [10 x 5]
##
##             batsman balls Played fours sixes  runs
##               (fctr)       (int) (dbl) (dbl) (dbl)
## 1       WU Tharanga          50     5     0    39
## 2        TM Dilshan          27     2     1    30
## 3     KC Sangakkara         103     4     1    75
## 4  DPMD Jayawardene          52     2     0    44
## 5     CK Kapugedera          17     0     0    17
## 6    TT Samaraweera           7     0     0     4
## 7       NLTC Perera           8     0     0     6
## 8        AD Mathews          22     1     1    36
## 9      HMRKB Herath           4     0     0     2
## 10       BAW Mendis           6     1     0     8
```

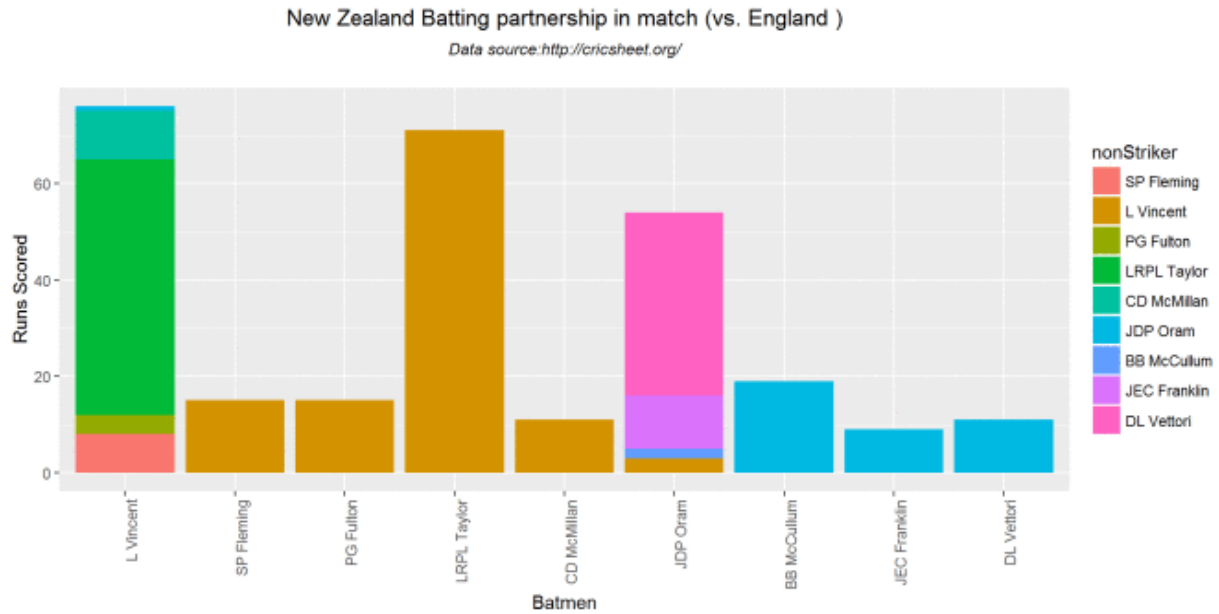
8. Plot the team batting partnerships

The functions below plot the team batting partnerships in the match **Note**: Many of the plots include an additional parameters plot which is either TRUE or FALSE. The default value is plot=TRUE. When plot=TRUE the plot will be displayed. When plot=FALSE the data frame will be returned to the user. The user can use this to create an interactive chary using one of the packages like rcharts, ggvis, googleVis or plotly.

```
teamBatsmenPartnershipMatch(pak_sa,"Pakistan","South Africa")
```



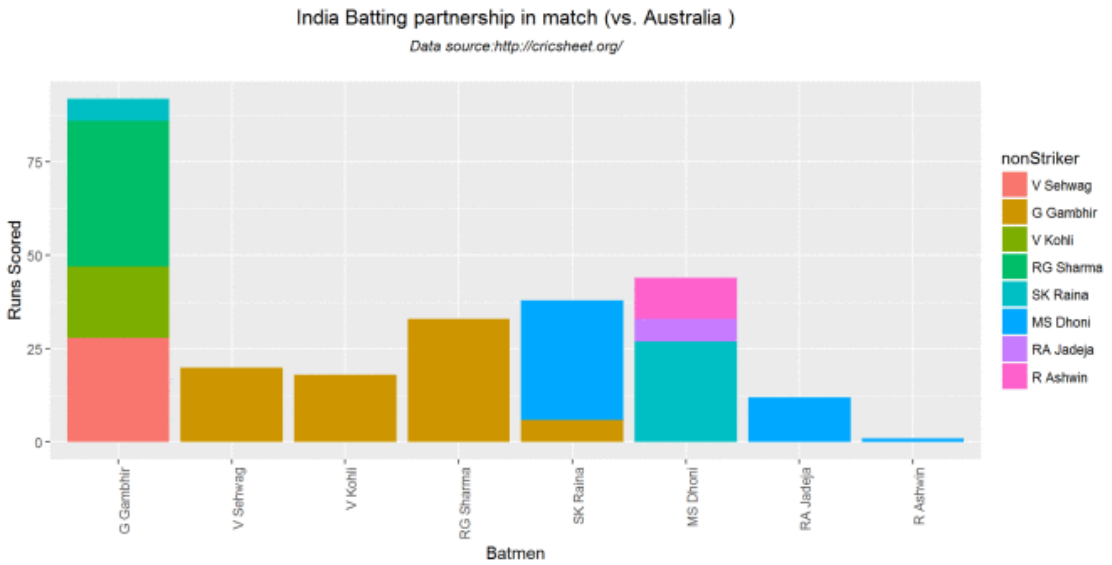

```
teamBatsmenPartnershipMatch(eng_nz,"New Zealand","England",plot=TRUE)
```




```
teamBatsmenPartnershipMatch(ban_zim,"Bangladesh","Zimbabwe",plot=FALSE)
##              batsman        nonStriker runs
## 1        Tamim Iqbal   Junaid Siddique    0
## 2        Tamim Iqbal Mohammad Ashraful    5
## 3    Junaid Siddique       Tamim Iqbal    0
## 4  Mohammad Ashraful       Tamim Iqbal    0
## 5  Mohammad Ashraful     Raqibul Hasan   20
## 6      Raqibul Hasan Mohammad Ashraful   13
```

```
## 7      Raqibul Hasan   Shakib Al Hasan    3
## 8     Shakib Al Hasan     Raqibul Hasan   12
## 9     Shakib Al Hasan   Mushfiqur Rahim    1
## 10   Mushfiqur Rahim   Shakib Al Hasan    1
## 11   Mushfiqur Rahim       Naeem Islam   30
## 12   Mushfiqur Rahim      Abdur Razzak    6
## 13   Mushfiqur Rahim      Dolar Mahmud   11
## 14   Mushfiqur Rahim     Rubel Hossain    8
## 15       Mahmudullah   Mushfiqur Rahim    4
## 16       Naeem Islam   Mushfiqur Rahim   21
## 17      Abdur Razzak   Mushfiqur Rahim    3

## 18      Dolar Mahmud   Mushfiqur Rahim   41
```

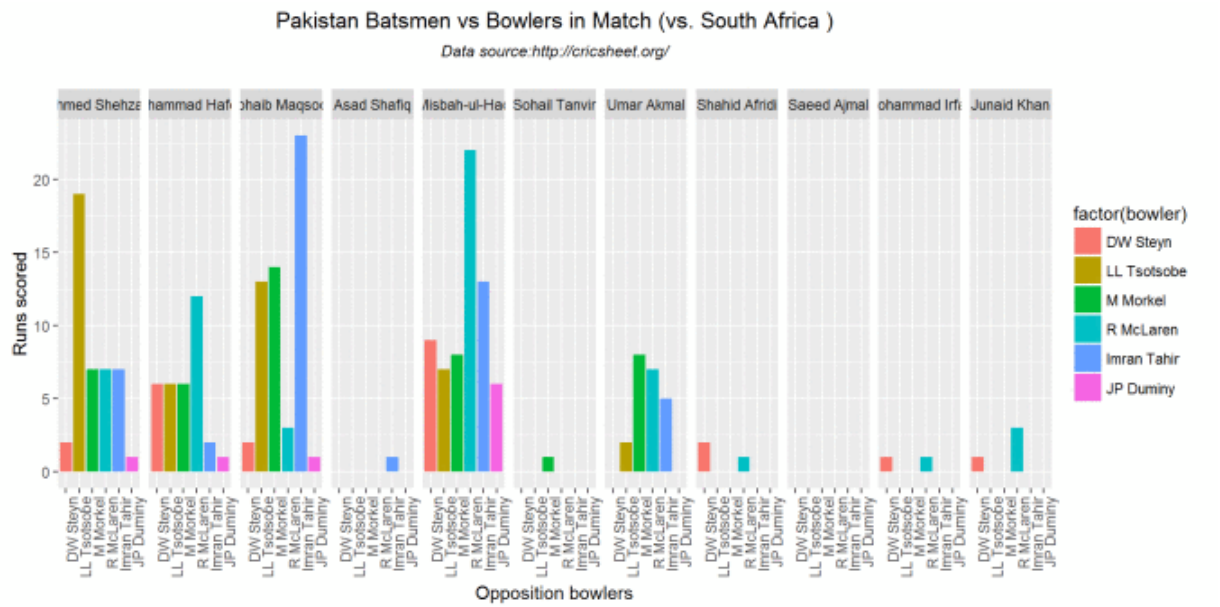
```
teamBatsmenPartnershipMatch(aus_ind,"India","Australia",  plot=TRUE)
```



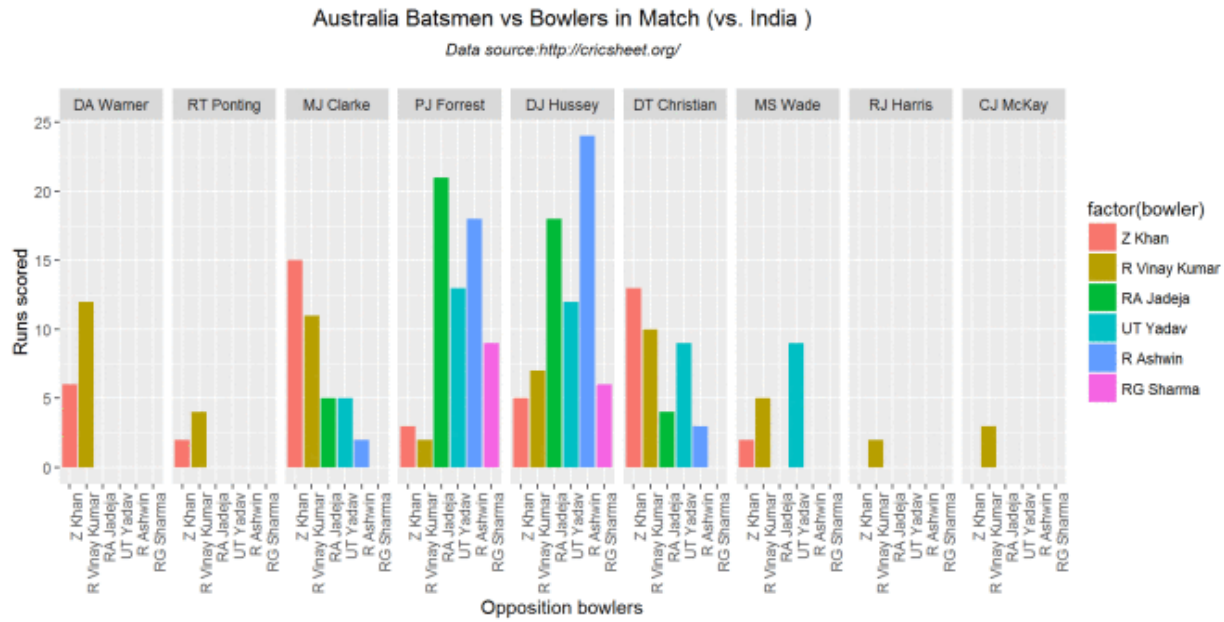
9. Batsmen vs Bowler

The function below computes and plots the performances of the batsmen vs the bowlers. As before the plot parameter can be set to TRUE or FALSE. By default it is plot=TRUE

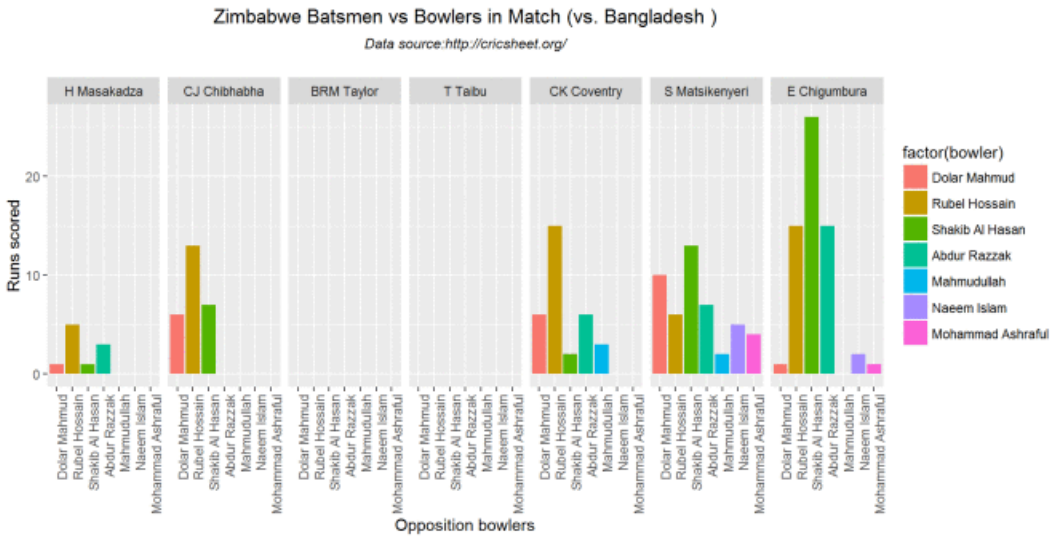
```
teamBatsmenVsBowlersMatch(pak_sa,'Pakistan',"South Africa",  plot=TRUE)
```




```
teamBatsmenVsBowlersMatch(aus_ind,'Australia',"India",plot=TRUE)
```




```
teamBatsmenVsBowlersMatch(ban_zim,'Zimbabwe',"Bangladesh",  plot=TRUE)
```




```
m <- teamBatsmenVsBowlersMatch(sl_wi,'West Indies',"Sri Lanka", plot=FALSE)

m
## Source: local data frame [35 x 3]
## Groups: batsman [?]
##
##      batsman        bowler runsConceded
##       (fctr)        (fctr)        (dbl)
## 1   CH Gayle  CRD Fernando            0
## 2   DM Bravo  CRD Fernando           15
## 3   DM Bravo   NLTC Perera           21
## 4   DM Bravo    AD Mathews           10
## 5   DM Bravo    BAW Mendis           11
## 6   DM Bravo CK Kapugedera            1
## 7   DM Bravo    TM Dilshan            5
## 8   DM Bravo  HMRKB Herath           16
## 9  AB Barath   NLTC Perera            0
## 10 RR Sarwan  CRD Fernando            6

## ..       ...           ...          ...
```

10. Bowling Scorecard

This function provides the bowling performance, the number of overs bowled, maidens, runs conceded and wickets taken for each match

```
teamBowlingScorecardMatch(eng_nz,'England')
## Source: local data frame [6 x 5]
##
##           bowler overs maidens  runs wickets
##           (fctr) (int)   (int) (dbl)   (dbl)
## 1    LE Plunkett     9       0    54       3
```

```
## 2    CT Tremlett    10       0    72       1
## 3     A Flintoff    10       0    66       0
## 4     MS Panesar    10       2    35       2
## 5  JWM Dalrymple     5       0    43       0

## 6 PD Collingwood     6       0    36       1
```

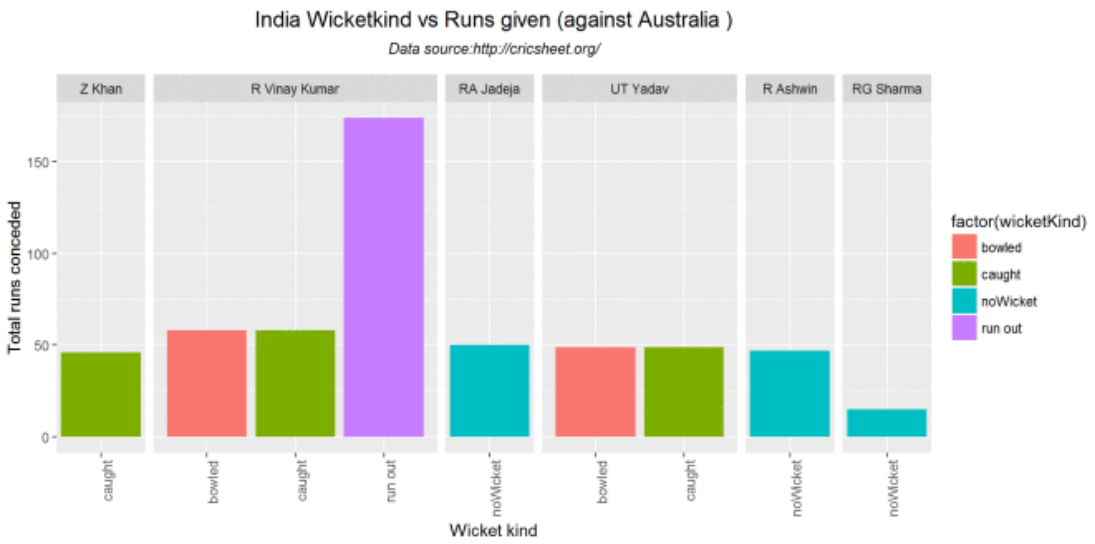
```
teamBowlingScorecardMatch(eng_nz,'New Zealand')
## Source: local data frame [6 x 5]
##
##         bowler overs maidens  runs wickets
##          (fctr) (int)   (int) (dbl)   (dbl)
## 1 JEC Franklin     8       1    45       1
## 2      SE Bond    10       0    58       1
## 3     JDP Oram     5       0    23       0
## 4     JS Patel    10       0    53       1
## 5   DL Vettori    10       0    40       3

## 6  CD McMillan     7       1    38       2
```

```
teamBowlingScorecardMatch(aus_ind,'Australia')
## Source: local data frame [6 x 5]
##
##         bowler overs maidens  runs wickets
##          (fctr) (int)   (int) (dbl)   (dbl)
## 1    RJ Harris    10       0    57       1
## 2     MA Starc     8       0    49       0
## 3     CJ McKay    10       1    53       3
## 4 DT Christian    10       0    45       0
## 5    DJ Hussey     3       0    13       0

## 6   XJ Doherty     9       0    51       2
```

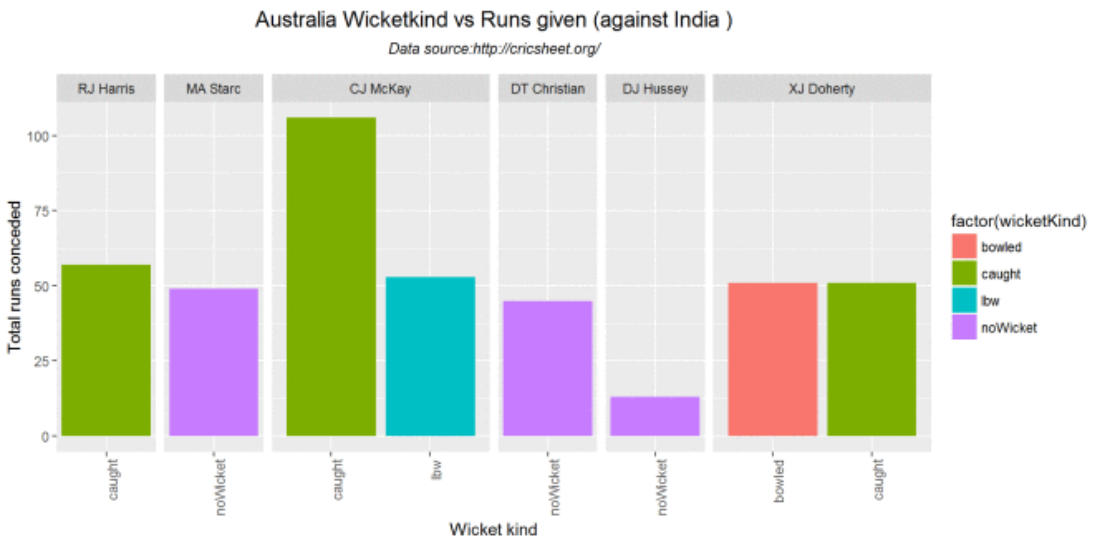
11. Wicket Kind

The plots below provide the bowling kind of wicket taken by the bowler (caught, bowled, lbw etc.)

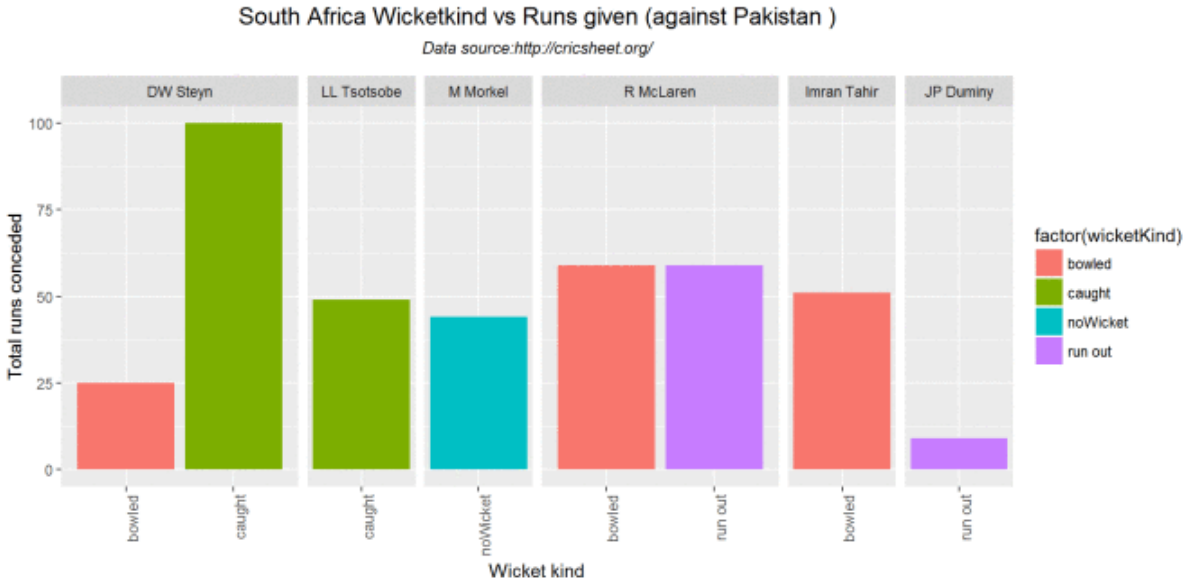
```
teamBowlingWicketKindMatch(aus_ind,"India","Australia")
```




```
teamBowlingWicketKindMatch(aus_ind,"Australia","India")
```




```
teamBowlingWicketKindMatch(pak_sa,"South Africa","Pakistan")
```

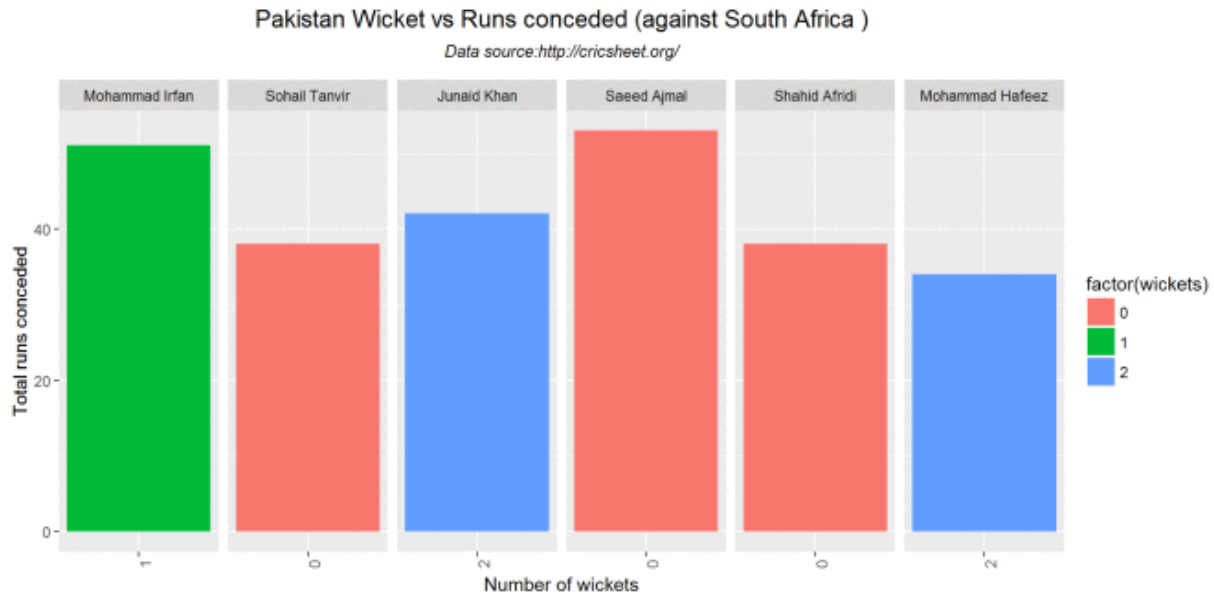



```
m  <-teamBowlingWicketKindMatch(sl_wi,"Sri Lanka",plot=FALSE)

m
##           bowler wicketKind wicketPlayerOut runs
## 1   CRD Fernando     bowled        CH Gayle   45
## 2    NLTC Perera     caught       AB Barath   36
## 3   HMRKB Herath        lbw       RR Sarwan   54
## 4     BAW Mendis     caught   S Chanderpaul   46
## 5    NLTC Perera        lbw        DM Bravo   36
## 6    NLTC Perera     caught       DJG Sammy   36
## 7   CRD Fernando     caught        DJ Bravo   45
## 8     BAW Mendis     caught       NO Miller   46
## 9     BAW Mendis     caught        CS Baugh   46
## 10    BAW Mendis     caught         SJ Benn   46
## 11    AD Mathews   noWicket        noWicket   33
## 12 CK Kapugedera   noWicket        noWicket    7

## 13    TM Dilshan   noWicket        noWicket   25
```

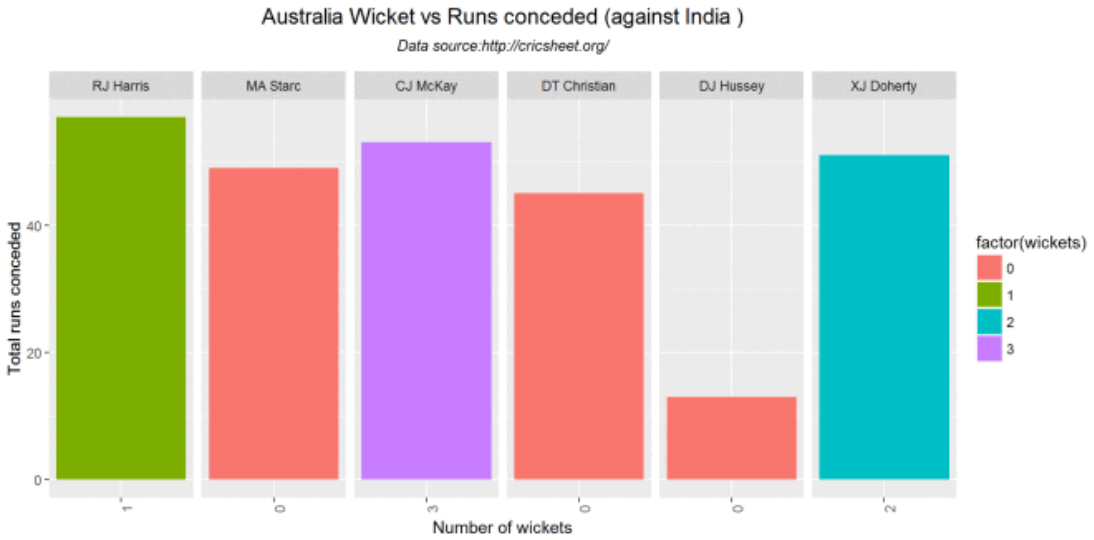
12. Wicket vs Runs conceded

The plots below provide the wickets taken and the runs conceded by the bowler in the match

```
teamBowlingWicketRunsMatch(pak_sa,"Pakistan","South Africa")
```




```
teamBowlingWicketRunsMatch(aus_ind,"Australia","India")
```




```
m <-teamBowlingWicketRunsMatch(sl_wi,"West Indies","Sri Lanka", plot=FALSE)

m
## Source: local data frame [6 x 5]
##
##       bowler overs maidens  runs wickets
##       (fctr) (int)   (int) (dbl)   (chr)
## 1 R Rampaul     5       0    44       1
## 2 DJG Sammy    10       1    61       1
## 3  DJ Bravo    10       0    58       3
## 4  CH Gayle    10       0    34       0
## 5   SJ Benn    10       1    38       4
```

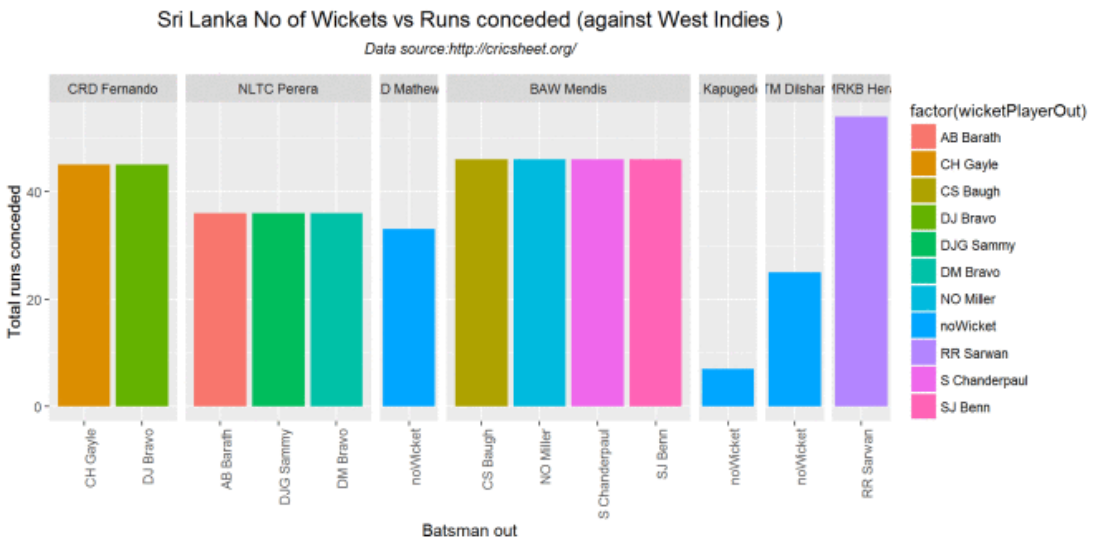
```
## 6 NO Miller      5        0     35        0
```

13. Wickets taken by bowler

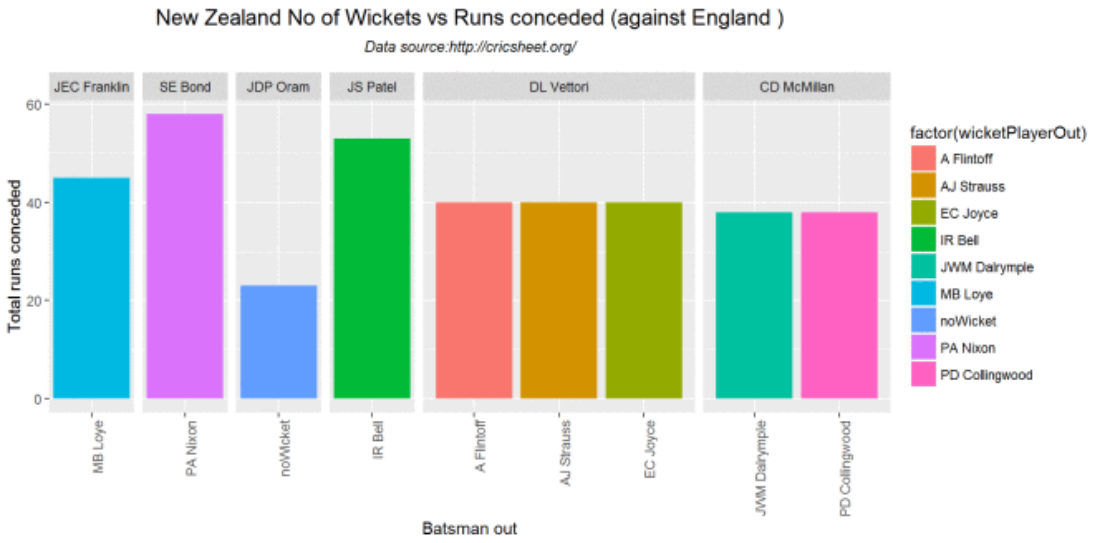
The plots provide the wickets taken by the bowler

```
m <-teamBowlingWicketMatch(eng_nz,'England',"New Zealand", plot=FALSE)

m
##            bowler wicketKind wicketPlayerOut runs
## 1     LE Plunkett        lbw       SP Fleming   54
## 2     LE Plunkett     caught        PG Fulton   54
## 3  PD Collingwood     caught     LRPL Taylor   36
## 4      MS Panesar    stumped     CD McMillan   35
## 5     LE Plunkett     caught        L Vincent   54
## 6      MS Panesar     caught     BB McCullum   35
## 7     CT Tremlett     caught    JEC Franklin   72
## 8      A Flintoff   noWicket         noWicket   66

## 9   JWM Dalrymple   noWicket         noWicket   43
```

```
teamBowlingWicketMatch(sl_wi,"Sri Lanka","West Indies")
```



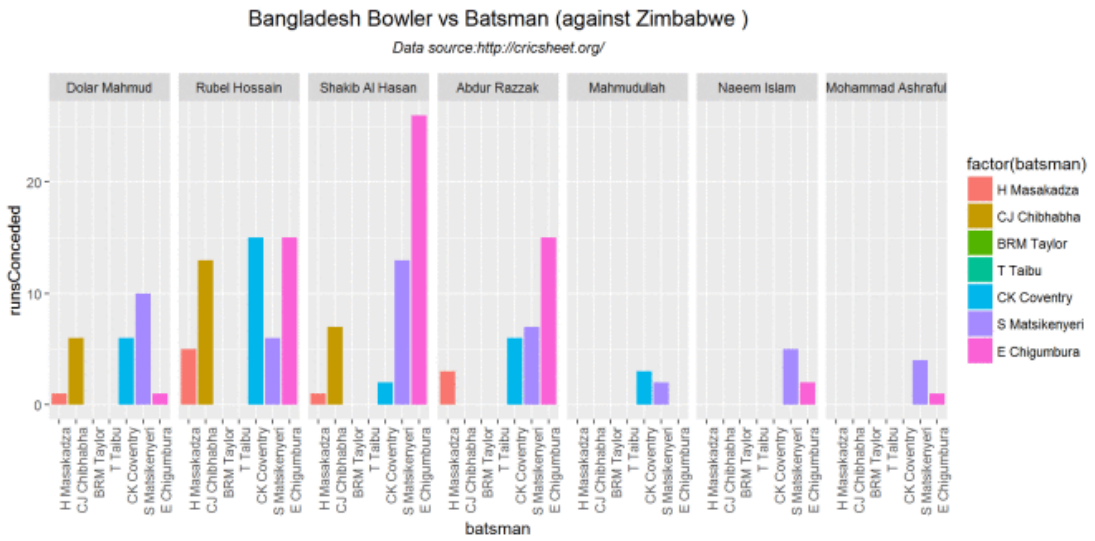

```
teamBowlingWicketMatch(eng_nz,"New Zealand","England")
```



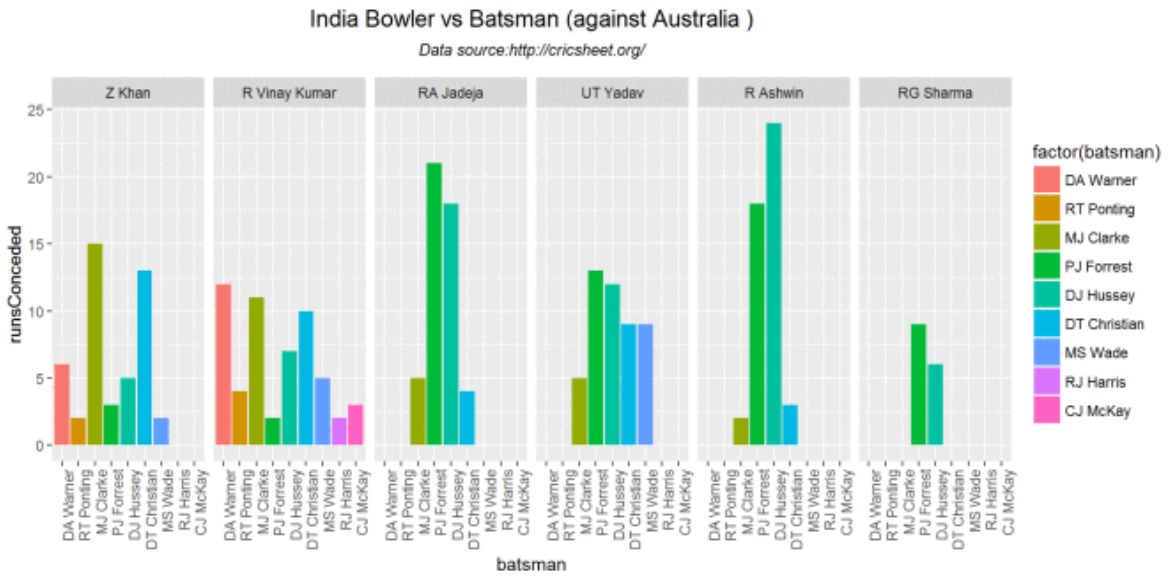
14. Bowler Vs Batsmen

The functions compute and display how the different bowlers of the country performed against the batting opposition.

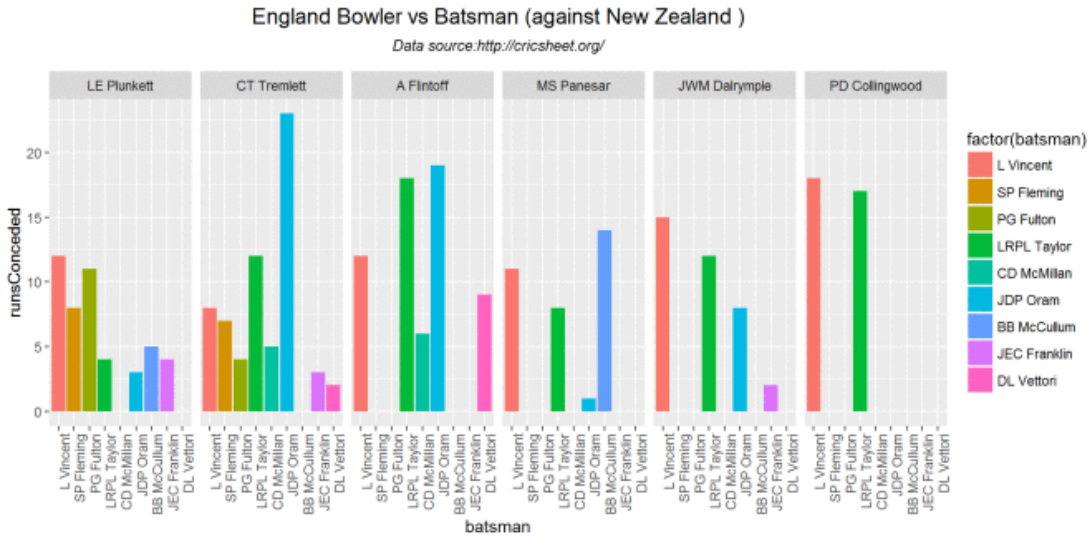
```
teamBowlersVsBatsmenMatch(ban_zim,"Bangladesh","Zimbabwe")
```




```
teamBowlersVsBatsmenMatch(aus_ind,"India","Australia")
```




```
teamBowlersVsBatsmenMatch(eng_nz,"England","New Zealand")
```



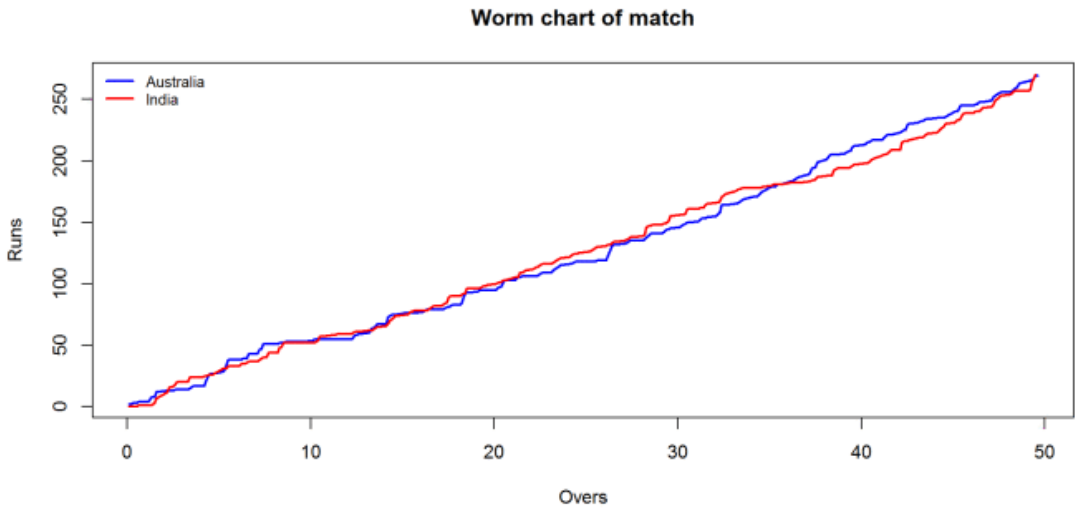

```
m <- teamBowlersVsBatsmenMatch(pak_sa,"Pakistan",plot=FALSE)

m
## Source: local data frame [30 x 3]
## Groups: bowler [?]
##
##            bowler        batsman runsConceded
##            (fctr)         (fctr)        (dbl)
## 1  Mohammad Irfan      Q de Kock           25
## 2  Mohammad Irfan        HM Amla           17
## 3  Mohammad Irfan   F du Plessis            0
## 4  Mohammad Irfan AB de Villiers            9
## 5   Sohail Tanvir      Q de Kock           11
```

```
## 6   Sohail Tanvir        HM Amla            6
## 7   Sohail Tanvir      JP Duminy            9
## 8   Sohail Tanvir      R McLaren           12
## 9     Junaid Khan      Q de Kock           24
## 10    Junaid Khan        HM Amla            6

## ..            ...            ...          ...
```

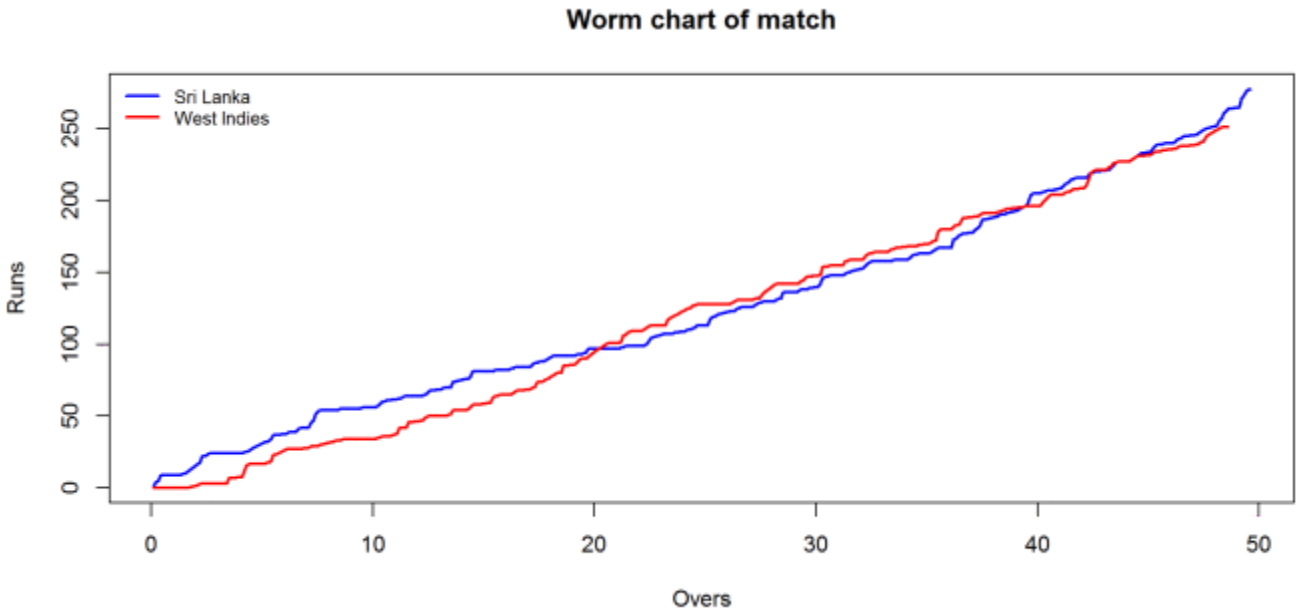
15. Match worm graph

The plots below provide the match worm graph for the matches

```
matchWormGraph(aus_ind,'Australia',"India")
```




```
matchWormGraph(sl_wi,'Sri Lanka',"West Indies")
```



Conclusion

This post included all functions between 2 opposing countries from the package yorkr.As mentioned above the yaml match files have been already converted to dataframes and are available for download from Github. Go ahead and give it a try!

About the author

Tinniam V Ganesh is a pioneer-architect-programmer with 28++ years' experience in software. He is a telecom veteran, cloud Architect and data scientist rolled in one. He is the author of 'cricketr'and 'yorkr' R packages, now available at CRAN.

Tinniam V Ganesh is an International IN patent holder (US6453028, EP11300928, JP3712624, CA2331958, KR2001008597).

He did his B-Tech in Electronics from IIT-BHU, Varanasi and Masters in Computer Science from Illinois Institute of Technology, Chicago. He has completed a 10 module Data Science Specialization from John Hopkins University, Bloomberg School of Public, besides also completing online courses on Machine Learning, Statistical Learning and Cloud Computing. He is currently working in IBM, Bangalore as a Cloud Architect in Cloud CoE. He can be contacted at tvganesh.85@gmail.com