ggplot2 Notes (Draft v1.01)

Vinu CT

September 18, 2012

A glimpse on Sachin's ODI scores

ggplot2 package

The main packages of R graphics are: base (Ross Ihaka), grid (Paul Murrell), lattics (Deepayan Sarkar), and ggplot2 (Hadley Wickham). This tutorial discuss about ggplot2. In my perspective, ggplot2 is the most elegant,capable, and high quality package in R. This package is an implementation of " The Grammar of Graphics, Wilkinson (2005)"

```
 ...In brief, the grammar tells us that a statistical graphic is a
 mapping  from data to aesthetic attributes (colour, shape, size)of
 geometric objects (points, lines, bars). The plot may also contain
 statistical transformations of the data and is drawn on a specific
coordinate  system...
                                                (from the ggplot2 book)
```

ggplot2 is developed by Hadley Wickham, assistant professor of statistics at Rice University, Houston. The latest version is 9.2, available in cran repositories.

This tutorial include mainly R codes for a case study related to cricket. The main stress of the case study is to explore ggplot2 package. The inference and artistic side of the graphs is left to you. Presentation and presentation slides cover some part of the "grammar of graphics" explanation. I have no doubt that understanding ggplot2 is worth investing your time.

I would be happy to get your suggestions. Please email me your valuable feedback at vinu.ct@hotmail.com.

About the cricket data

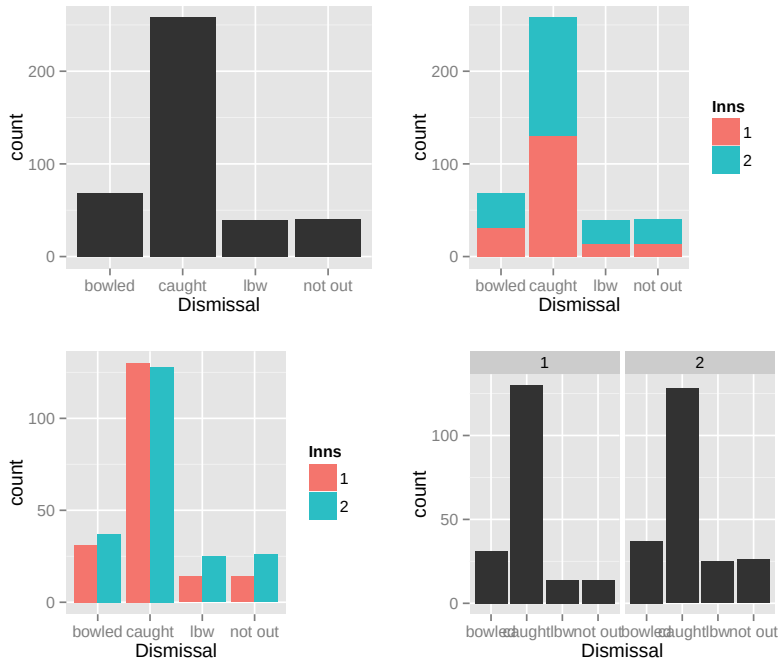
I was an occasional trainer when I was in corporate. The advantage I see there was most of the participants are familiar with the data. It was a fun to explore complex part of the data. But, here in academics (especially my institute), challenge in data analysis or R training session is to get familiar data set which involves certain complexity. Unfamiliarity of the dataset may disconnects the participants as the session progress. This motivated me to take a data from cricket. I have taken one day international (ODI) scores of Sachin Tendulkar (Considered to be one of the greatest batsmen of all time) from espn cricinfo website. Currently the analysis is restricted to Sachin's ODI data. I would see a detailed study on Indian team or all cricket data in future.

Tutorial

```
# install.package("ggplot2")  ## install & load ggplot library
library("ggplot2")
s10 <- read.csv(file="http://www.iimb.ernet.in/~vinuct10/ggplot/sachin.csv"
    )
```

```
 > tail(s10,3)
      X  StartDate Runs  BF     SR Pos Dismissal Inns   Opposition Ground result Score
450 450 2012-03-13    6  19  31.57   2     caught    1  v Sri Lanka  Dhaka    won 304/3
451 451 2012-03-16  114 147  77.55   2     caught    1 v Bangladesh  Dhaka   lost 289/5
452 452 2012-03-18   52  48 108.33   2     caught    2   v Pakistan  Dhaka    won 330/4
```

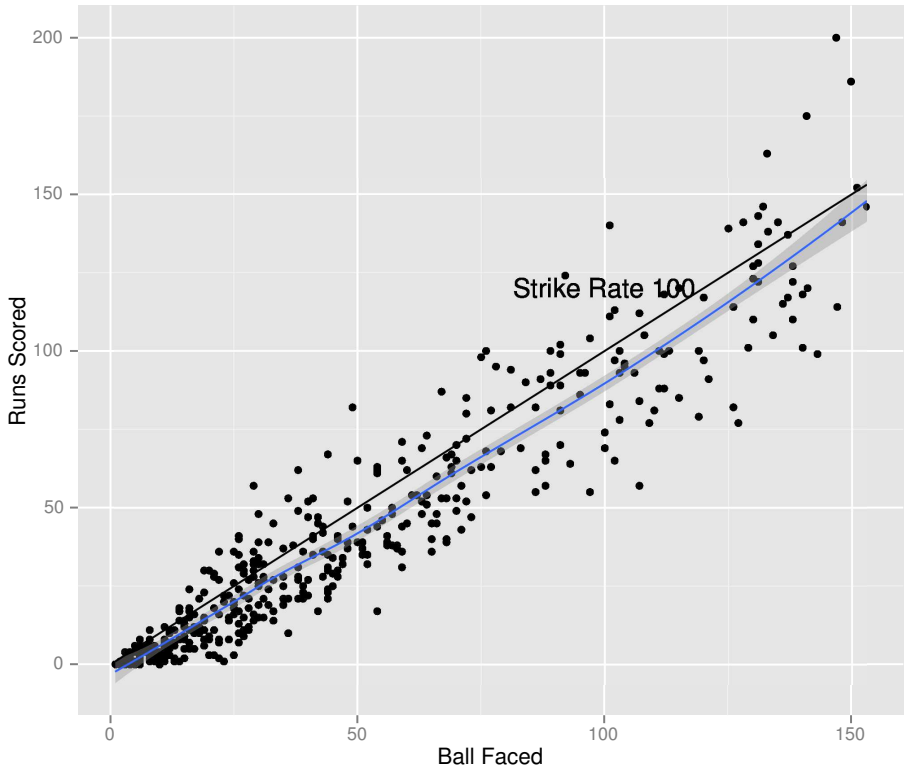
```
## Subsetting the dataset based on few dismissal
s10a <- subset(s10,Dismissal %in% c("bowled", "caught", "lbw","not out"))
s10a$Inns <- factor(s10a$Inns)

qplot( data=s10a, Dismissal, geom="bar")
qplot( data=s10a, Dismissal, geom="bar", fill=Inns)
qplot( data=s10a, Dismissal, geom="bar", fill=Inns,position="dodge")
qplot( data=s10a, Dismissal, geom="bar", facets=.~Inns )
```



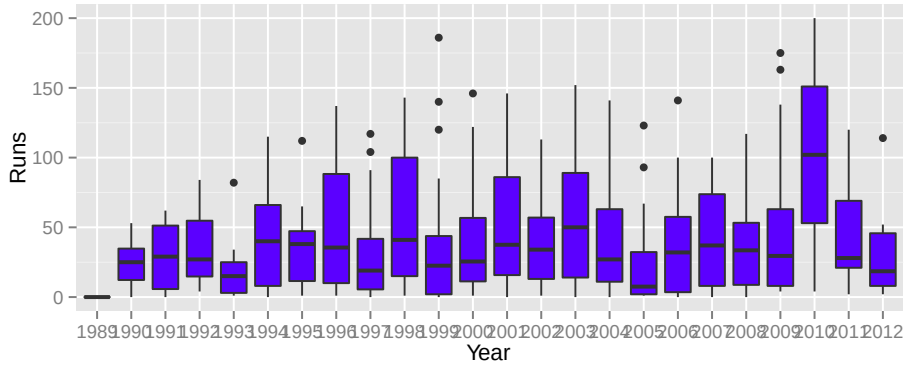
qplot (quick plot) is a convenient wrapper function for creating simple ggplot plot objects. The above codes can also write as

```
ggplot(data=s10a, aes(Dismissal)) + geom_bar()
ggplot(data=s10a, aes(Dismissal)) + geom_bar(aes(fill=Inns))
ggplot(data=s10a, aes(Dismissal)) + geom_bar(aes(fill=Inns,position="dodge"
    ))
ggplot(data=s10a, aes(Dismissal)) + geom_bar() + facet_grid(.~Inns)
```

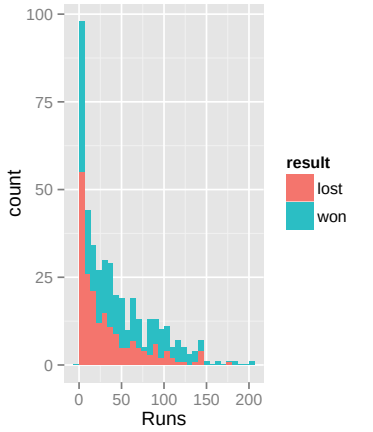
```
## Ball faced Vs Runs : scatter plot, splines, and SR 100 line
p1 <- ggplot(data=s10)+geom_point(aes(x=BF, y=Runs ))
p2 <- p1+geom_smooth(aes(x=BF, y=Runs )) +
         xlab("Ball Faced") + ylab("Runs Scored")
p3 <- p2+geom_line(data=s10,aes(BF,BF))+
         geom_text(aes(100, 120, label="Strike Rate 100"))
print(p3)
```

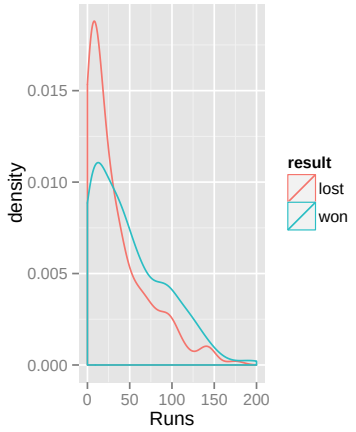


```
## Yearwise summary
require(chron)
qplot(x=factor(years(StartDate)), y=Runs, geom="boxplot", fill=I("blue"),
    data=s10)+
   xlab("Year")
```

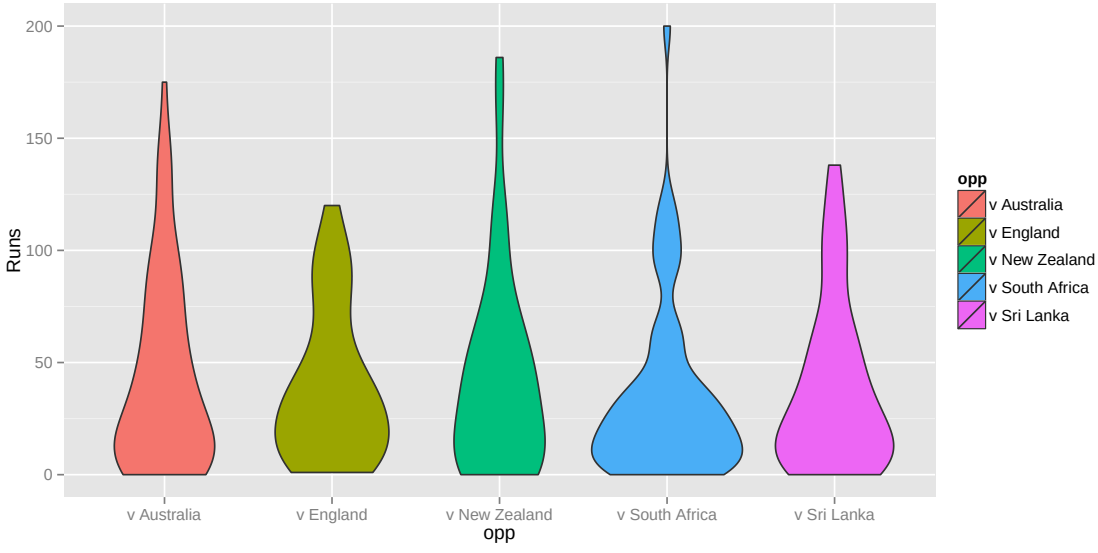


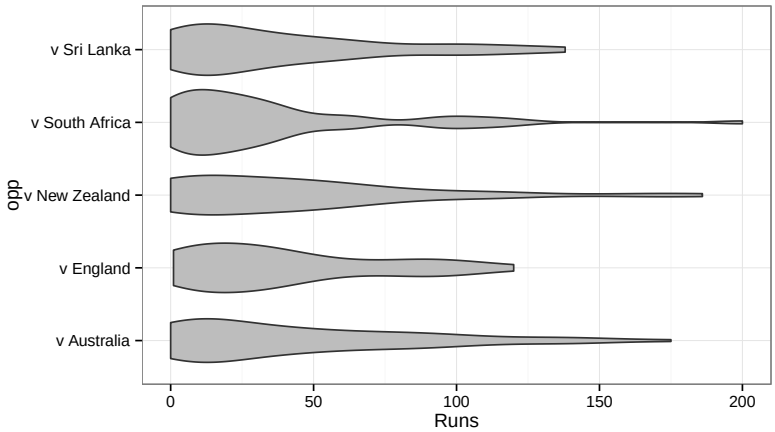

```
## Histogram of runs by match result
s10b <- subset(s10, result %in% c("won", "lost"))
s10b$result <- factor(s10b$result)

qplot(Runs, data=s10b, fill=result, geom=c("histogram"))
qplot(Runs, data=s10b, colour=result, geom=c("density"))
```



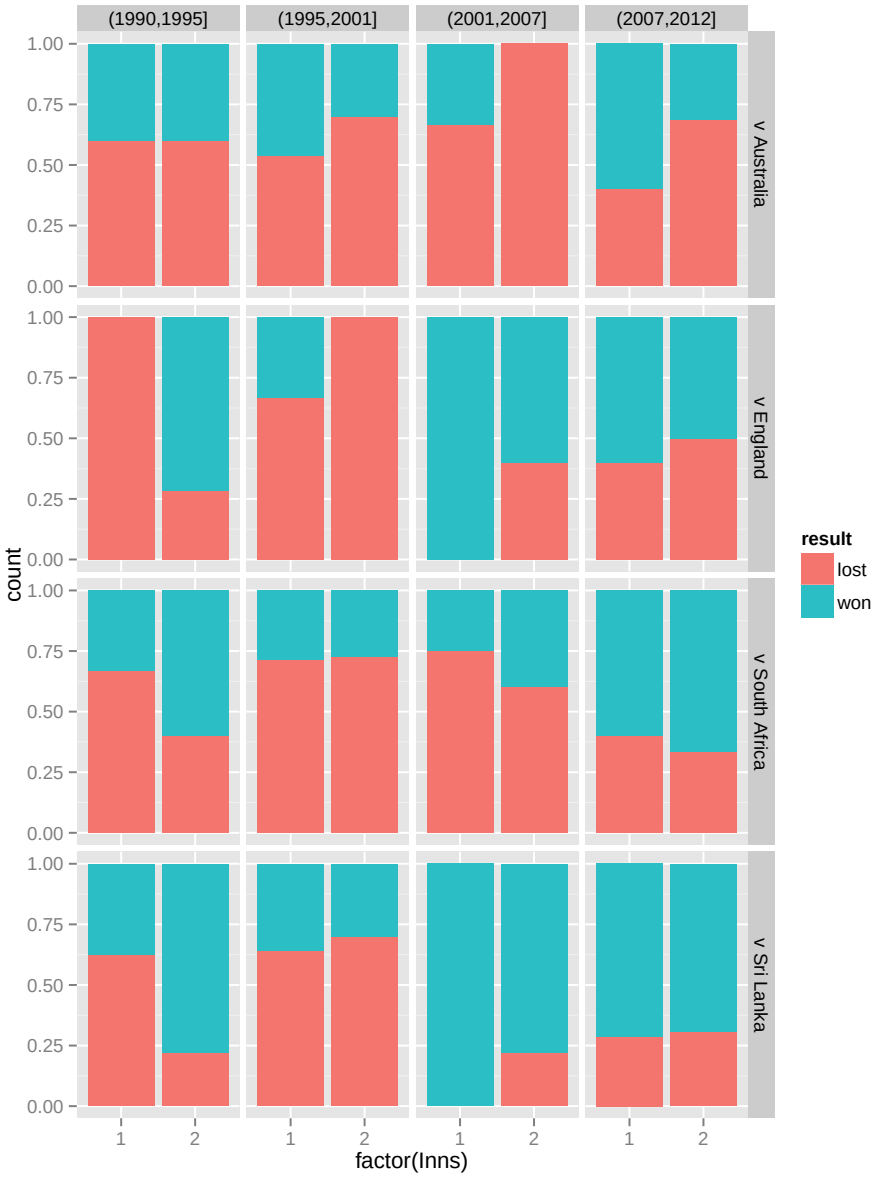


```
s10c <- subset(s10, Opposition %in% c("v Australia","v England","v South
    Africa","v New Zealand","v Sri Lanka"))
s10c$opp <- factor(s10c$Opposition)

## violin plots: Runs vs countries
 p1 <- ggplot(s10c, aes(opp, Runs))
p2 <- p1 + geom_violin(aes(fill=opp))
print(p2)
p3 <- p1 + geom_violin(fill=I("gray")) + coord_flip() + theme_bw()
print(p3)
```

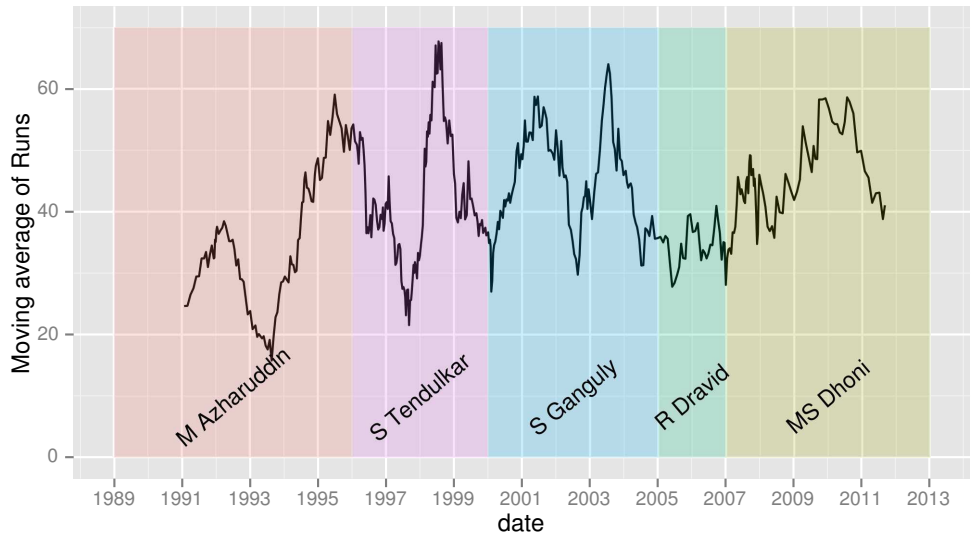




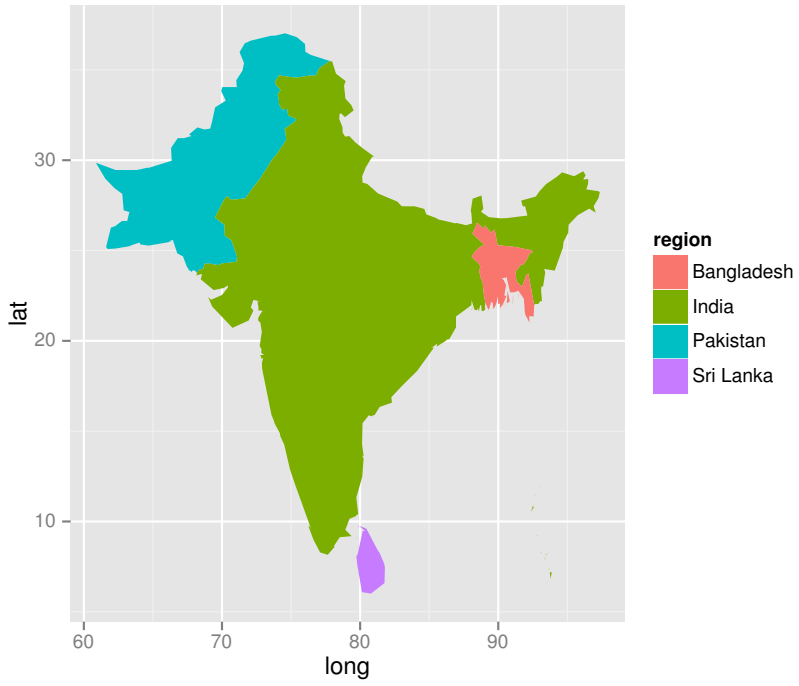
```
## Example of facets: Results of match by innings after faceting by
    Opposition team and Period
 s10c1= subset(s10,(Opposition %in% c("v Australia","v England","v South
     Africa","v Sri Lanka")) & (result %in% c("lost", "won"))  )
s10c1$yr= cut(as.numeric(format(s10c1$StartDate, format="%Y")),4)
qplot(data=s10c1, factor(Inns),fill=result,geom="bar",position="fill",
    facets= Opposition~yr)
```



```
## Sachin's performance over different captains
## The defintion of performance defined as follows
##  Performance at t = Average runs of 20 games around t (10 after and 10
    before)
require(zoo)
s10e<- cbind.data.frame(date=rollmean(s10$StartDate,20),avgRuns=rollmean(
    s10$Runs,20))

## Defining ODI captians history. The following information is created for
    demo purpose
start=as.Date(c("1989-01-01","1996-01-01","2000-01-01","2005-01-01","
    2007-01-01"))
end= as.Date(c("1996-01-01","2000-01-01","2005-01-01","2007-01-01","
    2013-01-01"))
mid=as.Date( (as.numeric(start) + as.numeric(end))/2)
captains <- cbind.data.frame(start,mid,end,captain=c("M Azharuddin","S
    Tendulkar","S Ganguly","R Dravid","MS Dhoni"))

p1 <- ggplot(aes(date,avgRuns),data=s10e)+ geom_line()+xlab("date")+ylab("
    Moving average of Runs") +ylim(c(0,70)) +
  scale_x_date(breaks="2 years", labels = date_format("%Y"))
p1+geom_rect(aes(NULL,NULL,xmin=start,xmax=end,fill=captain),ymin=0,ymax
    =70,alpha=0.2,data=captains) +
  geom_text(aes(x=mid,y=10,label=captain),data=captains,size=4,angle=40)+
     theme(legend.position="none")
```



```r
require(maps)
world <- map_data("world")
fasia <- world$region %in% c("India","Pakistan","Sri Lanka","Bangladesh")
asia <- world[fasia,]
qplot(long, lat, data = asia, geom = "polygon",fill=region, group = group)
```



TODO list

1. Few more graphs using facets, scales, and coordinate system.
2. Batting performance using survival analysis.
3. Customized plot inside cricket ground.
4. Documentation

Further Resources

- ggplot2 google group

- gglot2 cookbook
- Hadley Wickham's website
- Stack Overflow

Disclaimer: This is a draft version, circulated for comments. The case study included in this note is intended for discussion purposes only.