Data Visualization Using R

Sameer Bamnote

Cytel Statistical Software & Services Pvt. Ltd., India

July 19, 2014

Disclaimer

Any views or opinions presented in this presentation are solely those of the author and do not necessarily represent those of the company.

Outline

Introduction

Cytel

- Data visualization is the study of the visual representation of data.

Introduction

- Data visualization is the study of the visual representation of data.
- Presentation of data in a pictorial or graphical format.

Introduction

- Data visualization is the study of the visual representation of data.
- Presentation of data in a pictorial or graphical format.
- Data in spreadsheets is hard to visualize.

Introduction

Cytel

- Data visualization is the study of the visual representation of data.
- Presentation of data in a pictorial or graphical format.
- Data in spreadsheets is hard to visualize.
- It helps to see analytical results presented visually, find relevance among the millions of variables, communicate concepts and hypotheses to others, and even predict the future.

Introduction

Cytel

- Data visualization is the study of the visual representation of data.
- Presentation of data in a pictorial or graphical format.
- Data in spreadsheets is hard to visualize.
- It helps to see analytical results presented visually, find relevance among the millions of variables, communicate concepts and hypotheses to others, and even predict the future.
- Data visualization makes interpretation easier.

About R

- R is a opensource software.

About R

- R is a opensource software.
- Available for free and supported by strong user community.

About R

- R is a opensource software.
- Available for free and supported by strong user community.
- The current R is the result of a collaborative effort with contributions from all over the world.

About R

- R is a opensource software.
- Available for free and supported by strong user community.
- The current R is the result of a collaborative effort with contributions from all over the world.
- Developed by Ross Ihaka and Robert Gentelman at University of Auckland.

About R

Cytel

- R is a opensource software.
- Available for free and supported by strong user community.
- The current R is the result of a collaborative effort with contributions from all over the world.
- Developed by Ross Ihaka and Robert Gentelman at University of Auckland.
- Highly extensible and flexible.

About R

Cytel

- R is a opensource software.
- Available for free and supported by strong user community.
- The current R is the result of a collaborative effort with contributions from all over the world.
- Developed by Ross Ihaka and Robert Gentelman at University of Auckland.
- Highly extensible and flexible.
- Implementation of modern statistical methods.

Graphics in R

- R is a powerful environment for data visualization

Graphics in R

- R is a powerful environment for data visualization
- Integrated graphics

Graphics in R

- R is a powerful environment for data visualization
- Integrated graphics
- Good quality of graphics

Graphics in R

- R is a powerful environment for data visualization
- Integrated graphics
- Good quality of graphics
- Full control over graphics

Graphics in R

- R is a powerful environment for data visualization
- Integrated graphics
- Good quality of graphics
- Full control over graphics
- Can be reproduced

Graphics in R

- R is a powerful environment for data visualization
- Integrated graphics
- Good quality of graphics
- Full control over graphics
- Can be reproduced
- Huge number of R packages for graphics are available

Graphics in R

- R is a powerful environment for data visualization
- Integrated graphics
- Good quality of graphics
- Full control over graphics
- Can be reproduced
- Huge number of R packages for graphics are available
- R Graphs can be viewed on screen

Graphics in R

- R is a powerful environment for data visualization
- Integrated graphics
- Good quality of graphics
- Full control over graphics
- Can be reproduced
- Huge number of R packages for graphics are available
- R Graphs can be viewed on screen
- R graphs can be saved in various formats like pdf/png/jpg/wmf/tiff/ps etc.

Graphical Environments in R

Graphical Environments in R

Core - Overview

Cytel

Low-Level Graphics

Graphical Environments in R

Cytel

Low-Level Graphics

- Base graphics

Graphical Environments in R

Cytel

Low-Level Graphics

- Base graphics
- Grid graphics

Graphical Environments in R

Cytel

Low-Level Graphics

- Base graphics
- Grid graphics

High-Level Graphics

Graphical Environments in R

Cytel

Low-Level Graphics

- Base graphics
- Grid graphics

High-Level Graphics

- lattice

Graphical Environments in R

Cytel

Low-Level Graphics

- Base graphics
- Grid graphics

High-Level Graphics

- lattice
- ggplot2

Base graphics

Cytel

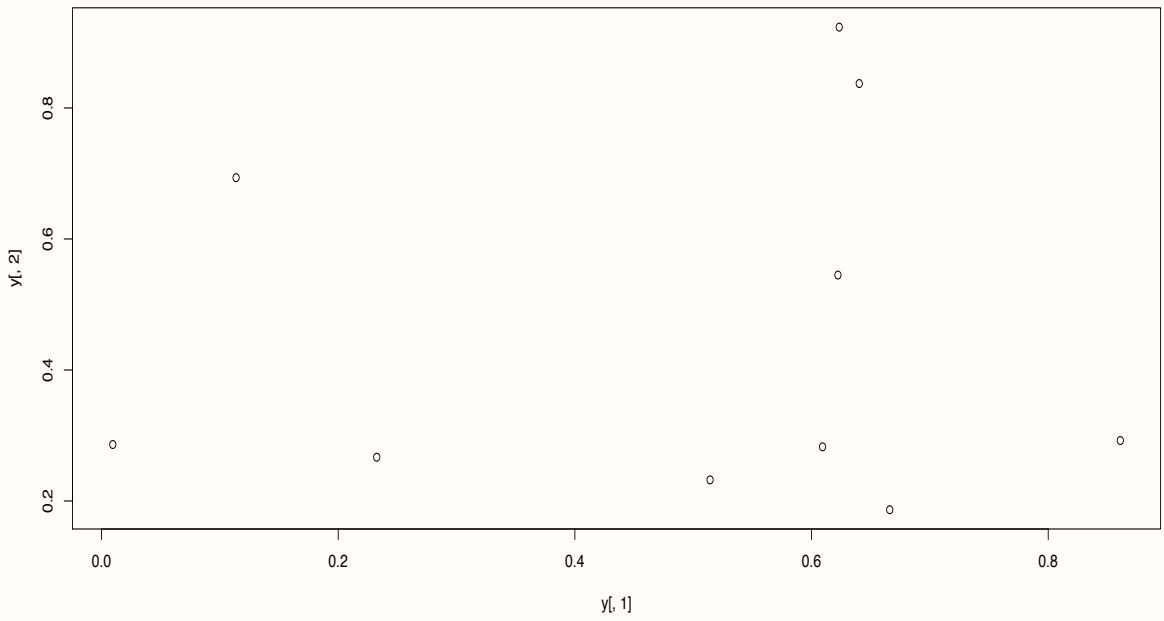
Plotting functions:

- plot - x-y plotting
- barplot - bar plots
- boxplot - box & whisker plot
- hist - histogram
- pie - pie charts
- dotchart - dot plots
- image, heatmap, contour, persp - functions to generate image like plots
- qqnorm, qqline, qqplot - distribution comparison plots
- pairs, coplot - display of multivariate data

Scatter plot: Basic

R code

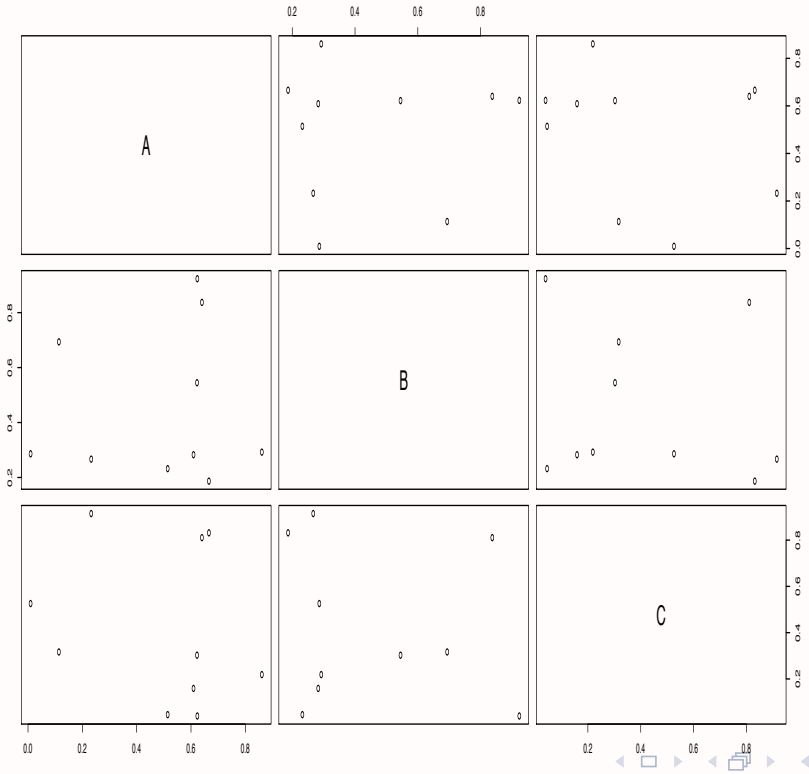
```
>  set.seed(1234)
> y <- matrix(runif(30), ncol=3, dimnames=list(letters[1:10], LETTERS[1:3]))
> plot(y[,1], y[,2])
```



Scatter plot: Pairs

R code

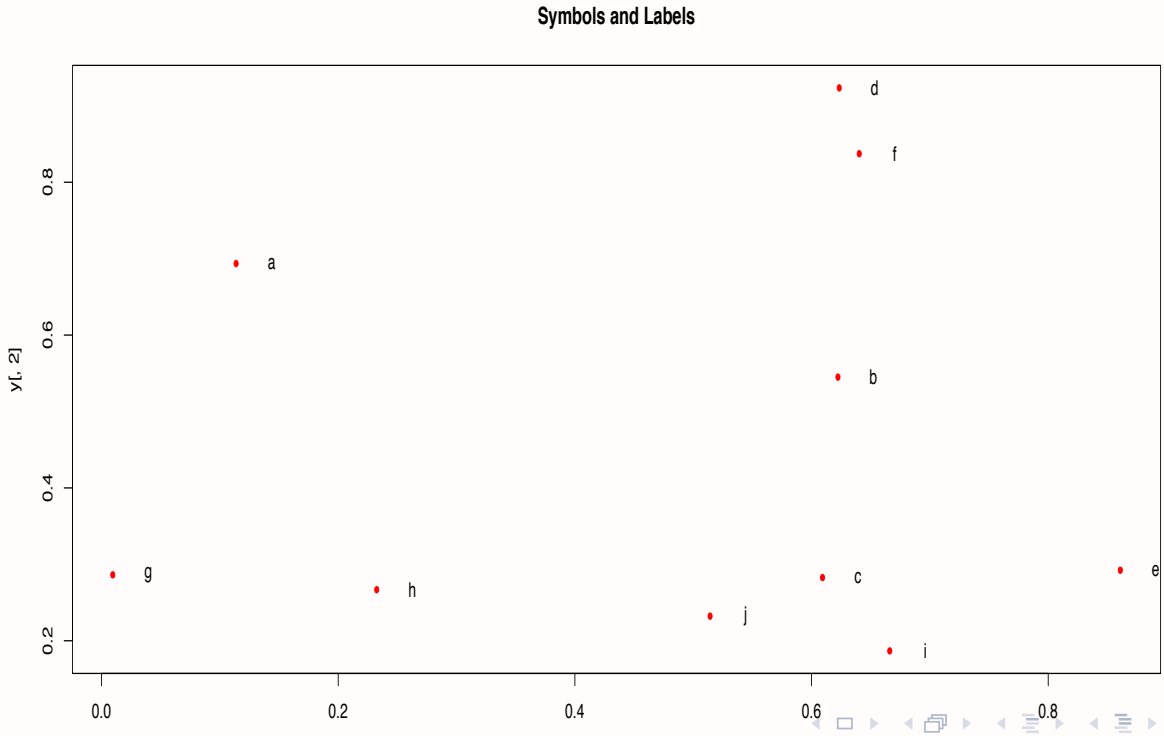
```
> pairs(y)
```



Scatter plot: with label

R code

```
> plot(y[,1], y[,2], pch=20, col="red", main="Symbols and Labels")
> text(y[,1]+0.03, y[,2], rownames(y))
```



Scatter plot: More Examples

Cytel

Using important plotting parameters

```
> grid(5, 5, lwd = 2)
> op <- par(mar=c(8,8,8,8), bg="lightblue")
> plot(y[,1], y[,2], type="p", col="red", cex.lab=1.2, cex.axis=1.2,
+ cex.main=1.2, cex.sub=1, lwd=4, pch=20, xlab="x label",
+ ylab="y label", main="My Main", sub="My Sub")
> par(op)
```

Used arguments

```
mar- species the margin sizes around the plotting area: c(bottom,left, top, right)
col- color of symbols
pch- type of symbols, samples: example(points)
lwd- size of symbols
cex.- control font sizes
```

Scatter plot: More examples

Adding a regression line

```
> plot(y[,1], y[,2])
> myline <- lm(y[,2]~y[,1]); abline(myline, lwd=2)
> summary(myline)
```

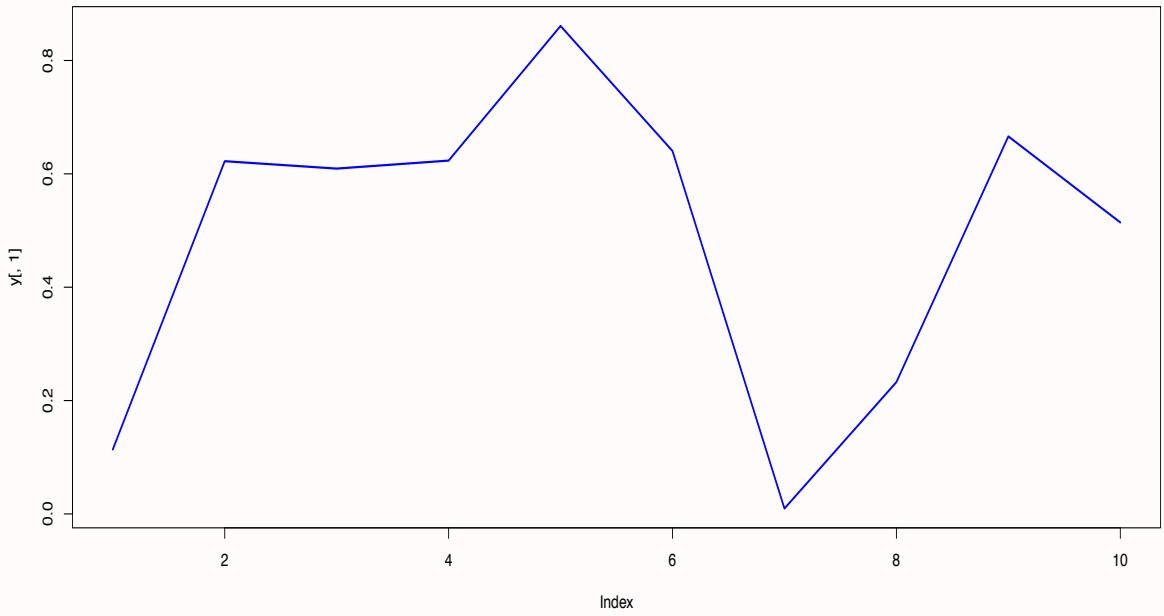
Plot on log scale

```
> plot(y[,1], y[,2], log="xy")
```

Line plot

R code

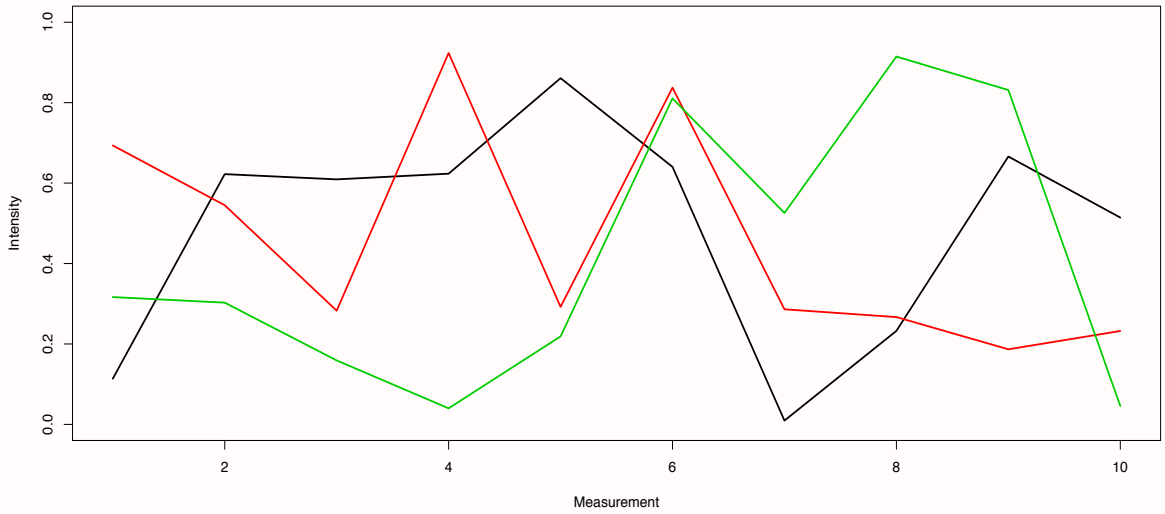
```
> plot(y[,1], type="l", lwd=2, col="blue")
```



Line plot with multiple variables

R code

```
> plot(y[,1], ylim=c(0,1),xlab="Measurement",ylab="Intensity",type="l",lwd=2,col=1)
> for(i in 2:length(y[1,])) {
+ screen(1, new=FALSE)
+ plot(y[,i], ylim=c(0,1), type="l", lwd=2, col=i, xaxt="n", yaxt="n", ylab="",
+ xlab="", main="", bty="n")
+ }
```



More Functions in Base graphics

Cytel

Graph Type	R function

More Functions in Base graphics

Cytel

Graph Type	R function
• Barplot	• barplot

More Functions in Base graphics

Cytel

Graph Type	R function
Barplot	barplot
Histogram	hist

More Functions in Base graphics

Cytel

Graph Type	R function
Barplot	barplot
Histogram	hist
Density plot	density

More Functions in Base graphics

Graph Type	R function
Barplot	barplot
Histogram	hist
Density plot	density
Pie Chart	pie

Grid Graphics Environment

Cytel

What is grid?

- Low-level graphics system
- Highly exible and controllable system
- Does not provide high-level functions
- Intended as development environment for custom plotting functions
- Pre-installed on new R distributions

Lattice

Cytel

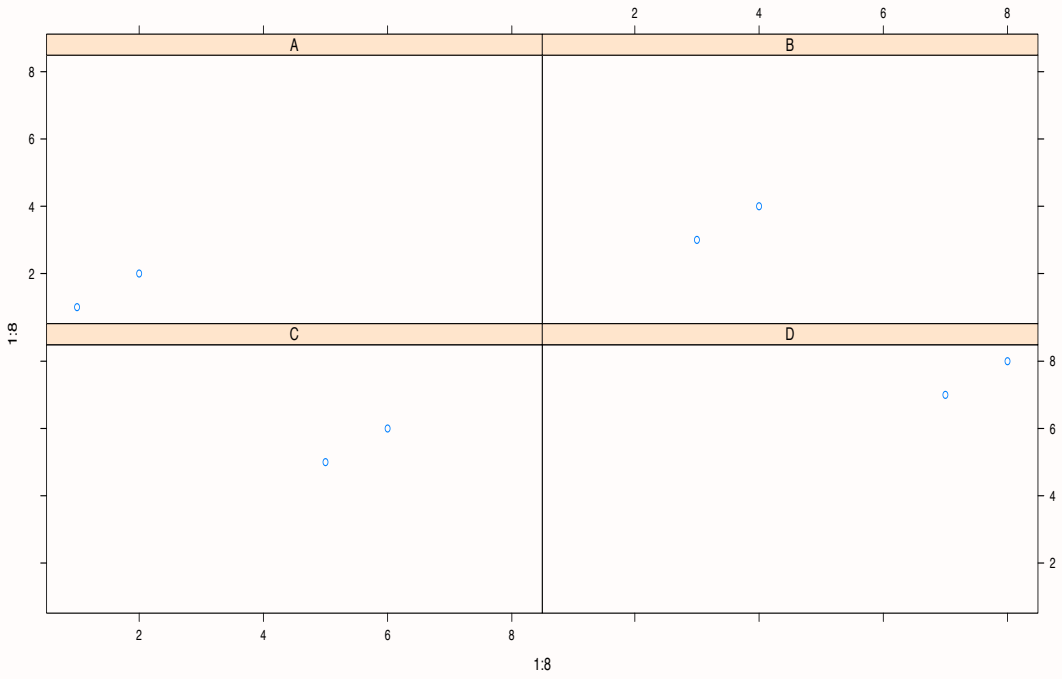
What is lattice?

- High-level graphics system
- Developed by Deepayan Sarkar
- Implements Trellis graphics system from S-Plus
- Simplies high-level plotting tasks, arranging complex graphical features
- Syntax similar to R's base graphics

Lattice plot example

R code

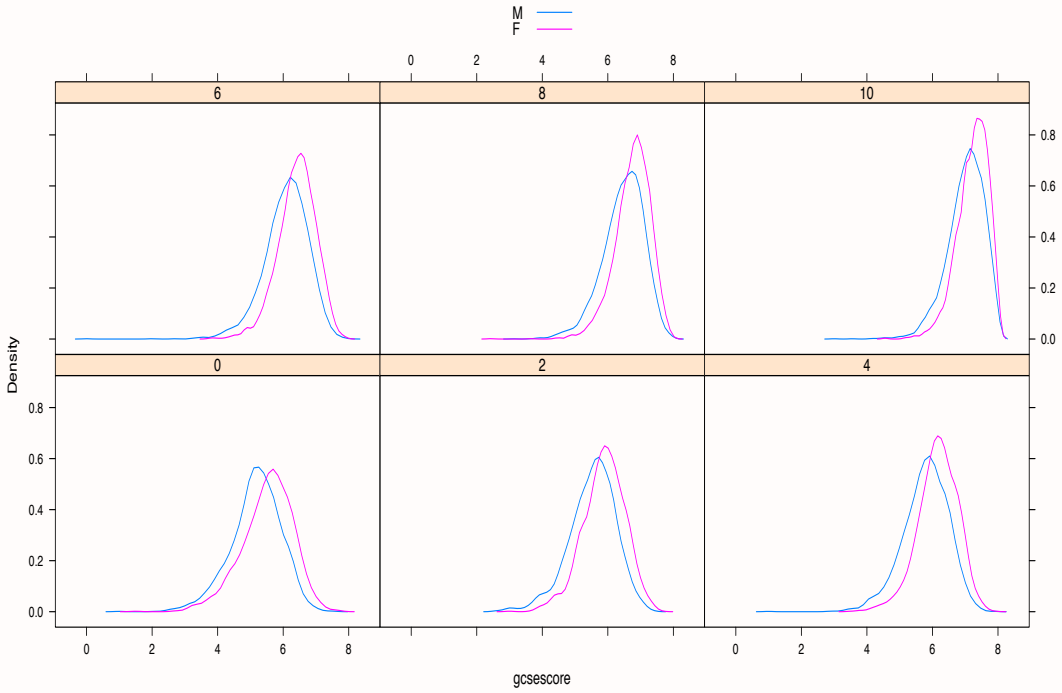
```
> library(lattice)
> p1 <- xyplot(1:8 ~ 1:8 | rep(LETTERS[1:4], each=2), as.table=TRUE)
> plot(p1)
```



Lattice plot: Density plot

R code

```
> data(Chem97, package = "mlmRev")
> densityplot(~ gcsescore | factor(score),data= Chem97, groups = gender,
+ plot.points = FALSE, auto.key = TRUE)
```



ggplot2

Cytel

What is ggplot2?

- High-level graphics system
- Implements grammar of graphics from Leland Wilkinson
- Streamlines many graphics workows for complex plots
- Syntax centered around main ggplot function
- Simpler qplot function provides many shortcuts
- ggplot function accepts two arguments (1)Data set to be plotted (2)Aesthetic mappings provided by aes function

qplot function

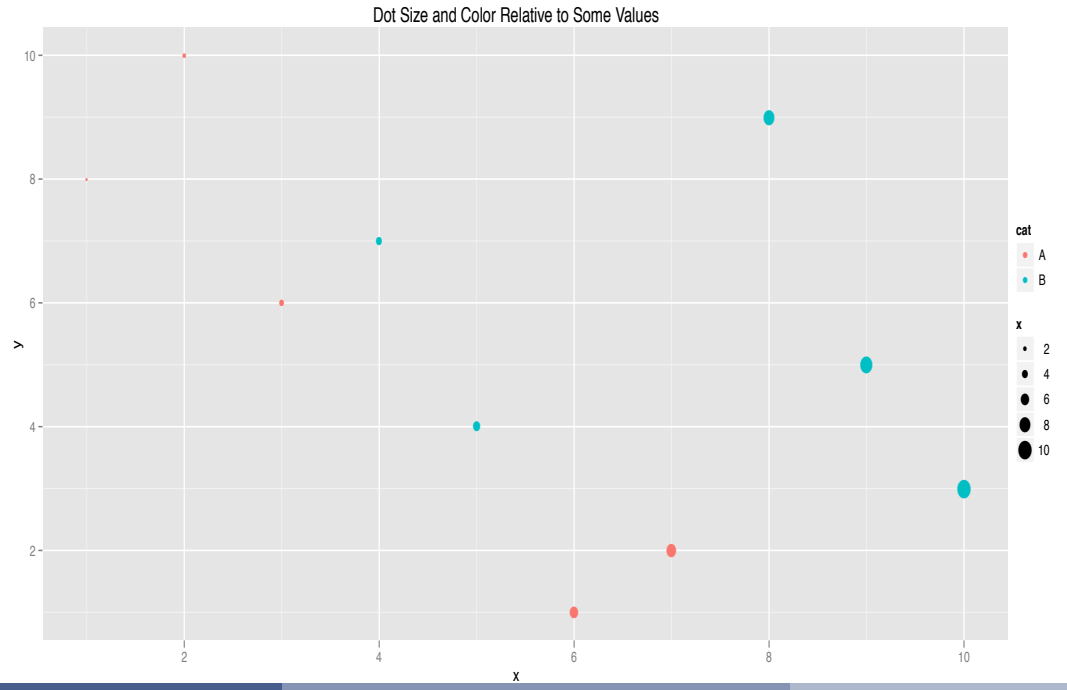
Cytel

- qplot syntax is similar to R's basic plot function
- Arguments:
  - x: x-coordinates (e.g. col1)
  - y: y-coordinates (e.g. col2)
  - data: data frame with corresponding column names
  - xlim, ylim: e.g. xlim=c(0,10)
  - log: e.g. log="x" or log="xy"
  - main: main title; see ?plotmath for mathematical formula
  - xlab, ylab: labels for the x- and y-axes
  - color, shape, size
  - ...: many arguments accepted by plot function

qplot: Scatterplot

R code

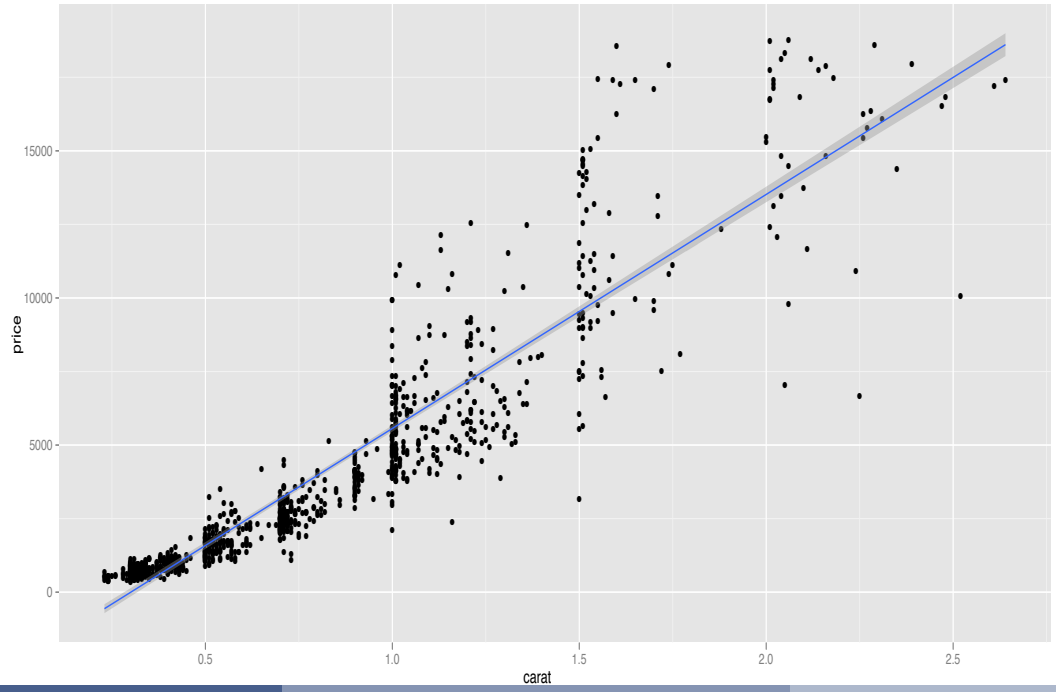
```
> p <- qplot(x, y, geom="point", size=x, color=cat,
+ main="Dot Size and Color Relative to Some Values") +
+ theme(legend.position = "none")
> print(p)
```



qplot:Scatterplot with Regression Line

R code

```
> set.seed(1234)
> dsmall <- diamonds[sample(nrow(diamonds), 1000), ]
> p <- qplot(carat, price,data = dsmall,geom = c("point", "smooth"),method= "lm")
> print(p)
```



ggplot: Scatterplot

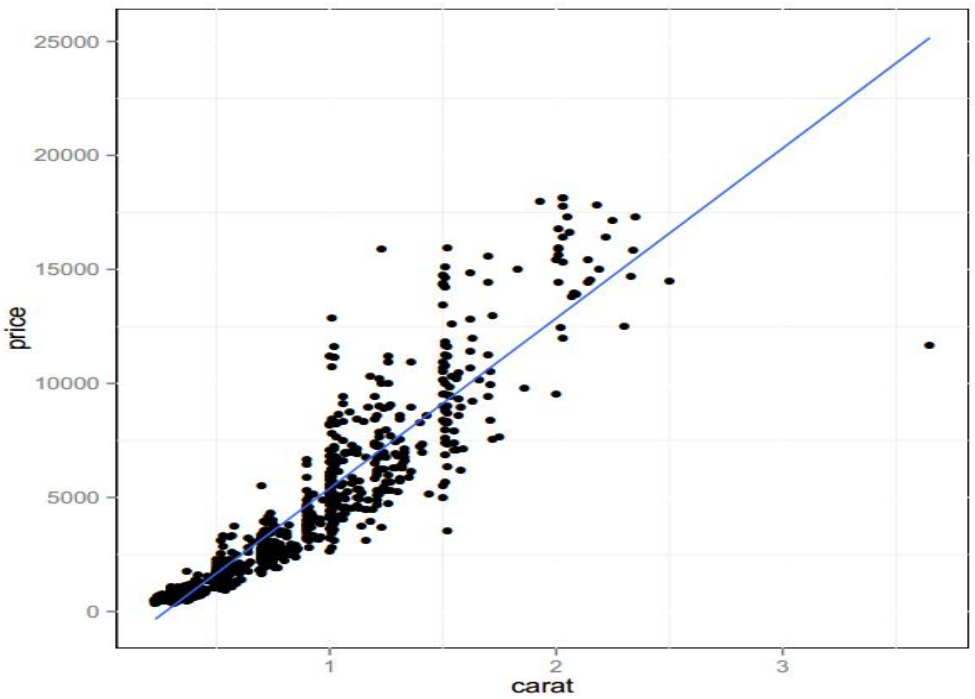
R code

```
> p <- ggplot(dsmall, aes(carat, price, color=color)) + geom_point(size=4)
> print(p)
```

ggplot: Scatterplot with Regression Line

R code

```
> p <- ggplot(dsmall, aes(carat, price, color=color)) + geom_point(size=4)
> print(p)
```



ggplot: Jitter Plot

R code

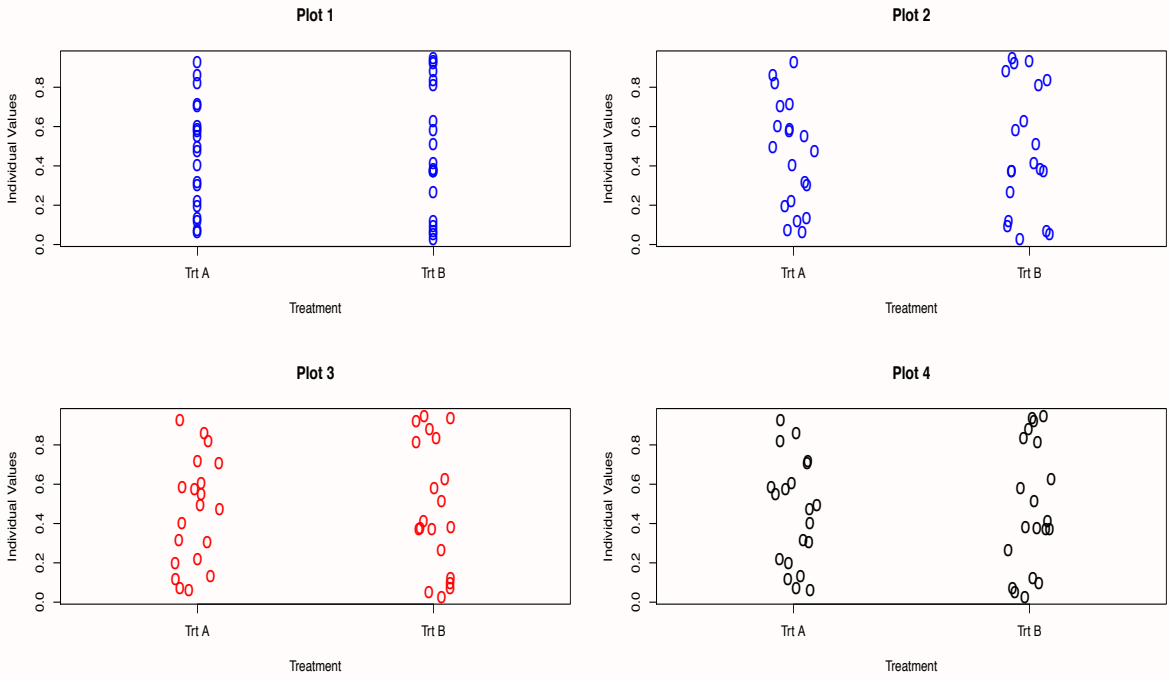
```
> p <- ggplot(dsmall, aes(color, price/carat)) +
+ geom_jitter(alpha = I(1 / 2), aes(color=color))
>  print(p)
```

Multiple Figures in Single Panel

- Requirement - Multiple figures in a single panel

Multiple Figures in Single Panel

- Requirement - Multiple figures in a single panel



Multiple graphs in Single Panel

R Code

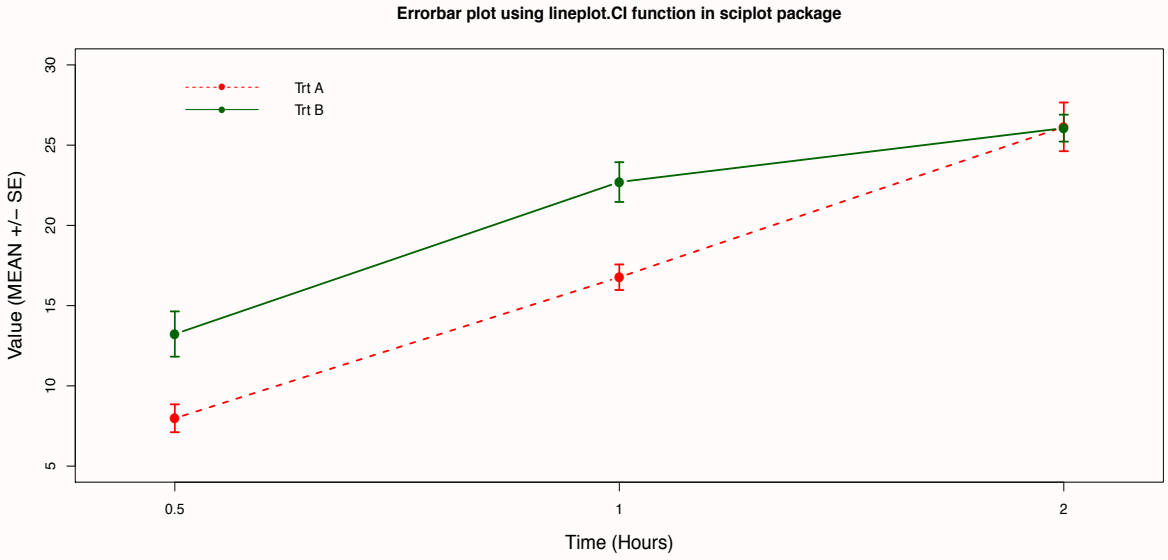
```
pdf("D:/Example1.pdf")
par(mfrow=c(2,2)) # Divide the plotting area

#<Code for Plot 1>
plot(b,a, xaxt = 'n', xlab = 'Treatment',ylab = 'Individual Values',
xlim=c(1,5),main = "Plot 1",cex=1.5,col='blue', pch = 'o')
axis(1,at=c(2,4),labels = c('Trt A', 'Trt B'))
#<Code for Plot 2>
plot(jitter(b,amount=0.2),a, xaxt = 'n', xlab = 'Treatment',ylab =
'Individual Values', xlim=c(1,5),main = "Plot 2", cex=1.5,col='blue',
pch = 'o')
axis(1,at=c(2,4),labels = c('Trt A', 'Trt B'))
# <Code for Plot 3>
# <Code for Plot 4>

dev.off()
```

Errorbar Plot

- Using an inbuilt data in R named 'ToothGrowth' whose variables are renamed as rep, trt and time.
- Errorbar plot using 'lineplot.CI' from 'sciplot' package in R



Errorbar plot using sciplot package

R Code

```
lineplot.CI(time, rep, group = trt, data = tg, cex = 1.5,
  xlab = "Time (Hours)", ylab = "Value (MEAN +/- SE)",
  cex.lab = 1.3, x.leg = 1,y.leg=30,col = c("red","dark green"),
  pch = c(16,16),ylim=c(5,30), err.width = 0.05, xaxt = 'n',lwd=2)

axis(1,at=c(1,2,3),labels=c(0.5, 1, 2))

title("Errorbar plot using 'lineplot.CI' function in sciplot package")
```

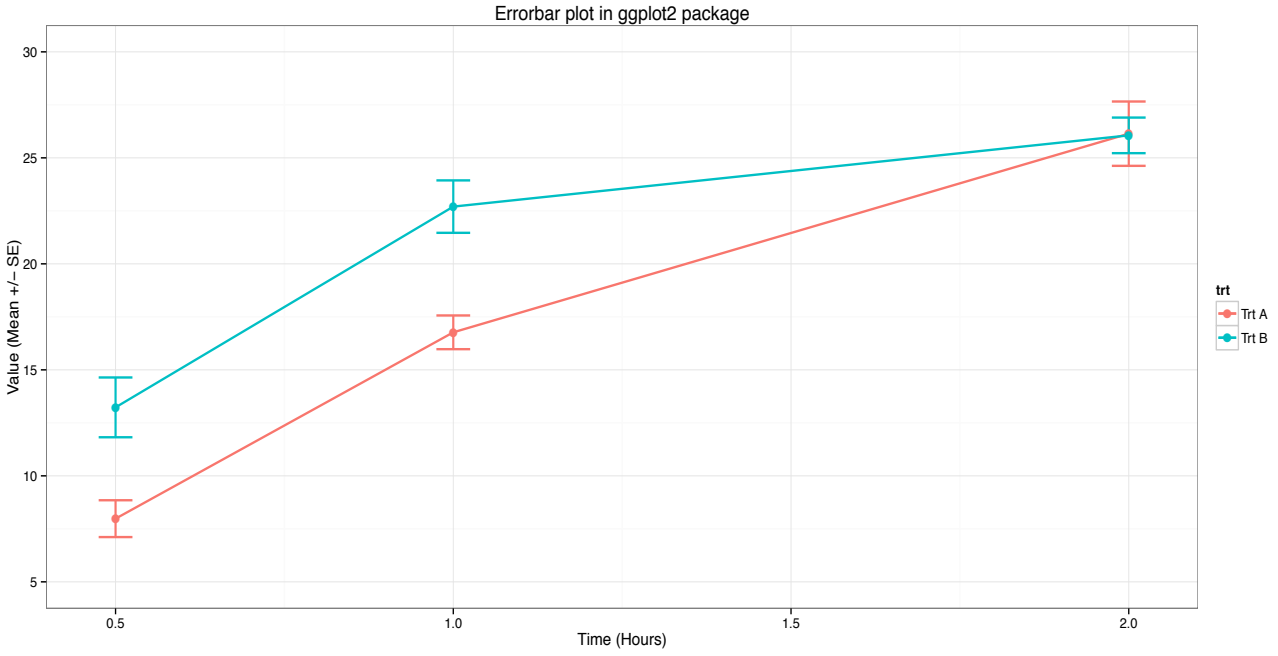
Errorbar plot using ggplot2 package

Cytel

- Errorbar plot using 'ggplot2' package in R

Errorbar plot using ggplot2 package

- Errorbar plot using 'ggplot2' package in R



Errorbar plot using ggplot2 package

R Code

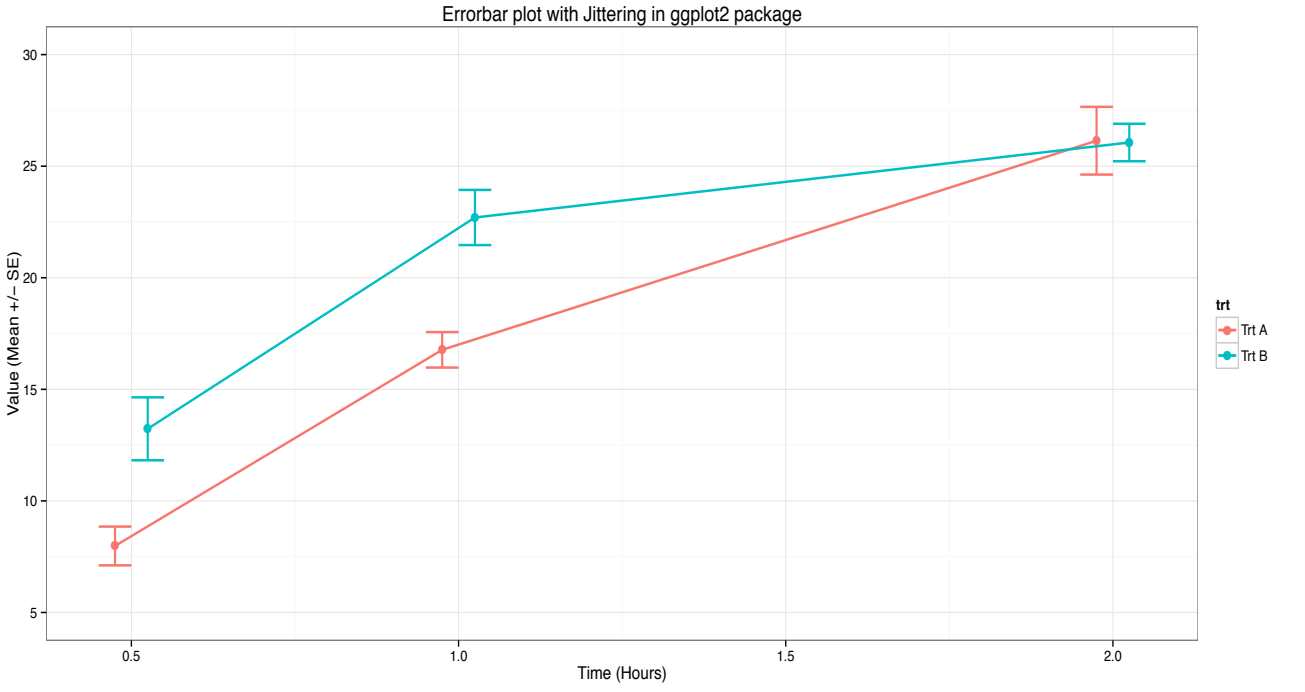
```
ggplot(summary, aes(x=time, y=rep, colour=trt)) +
 geom_errorbar(aes(ymin=rep-se, ymax=rep+se), width=.05, lwd = 0.8) +
 geom_line(lwd=0.8) +
 geom_point(cex=3) + xlab("Time (Hours)") + ylab ("Value (Mean +/- SE)")+
 ggtitle("Errorbar plot in ggplot2 package")+
 theme_bw()+
 scale_y_continuous(limits=c(5,30), breaks=0:30*5)
```

Errorbar plot using ggplot2 package

R Code

```
pd <- position_dodge(.1)
ggplot(summary, aes(x=time, y=rep, colour=trt)) +
 geom_errorbar(aes(ymin=rep-se, ymax=rep+se), width=.1, position=pd,
 lwd =0.8) + geom_line(position=pd,lwd =0.8) + geom_point(position=pd,
 cex = 3) + xlab ("Time (Hours)") + ylab ("Value (Mean +/- SE)")+ ...
```

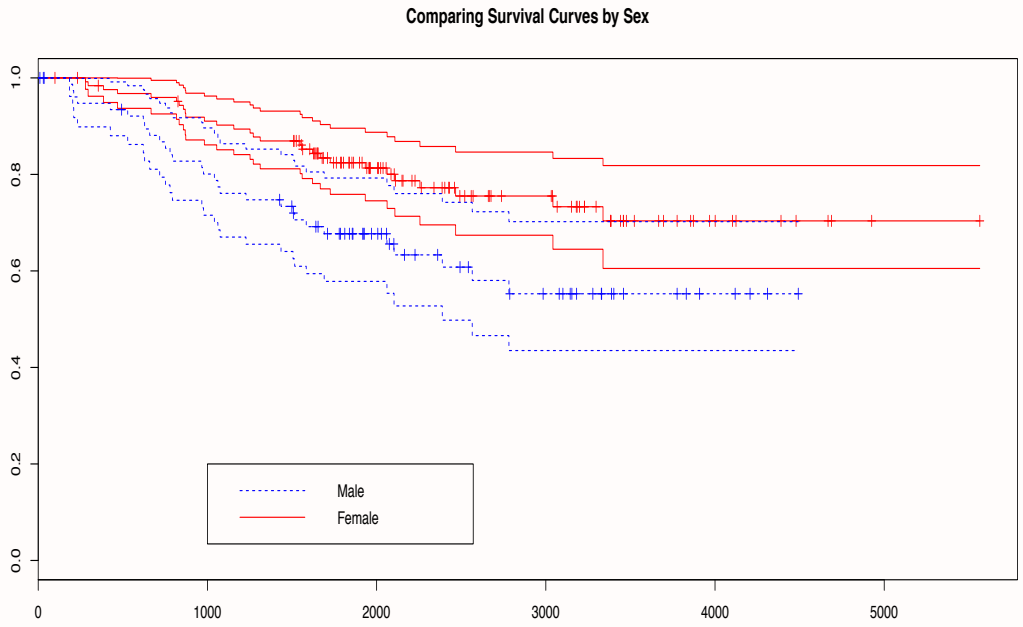
Errorbar plot using ggplot2 package



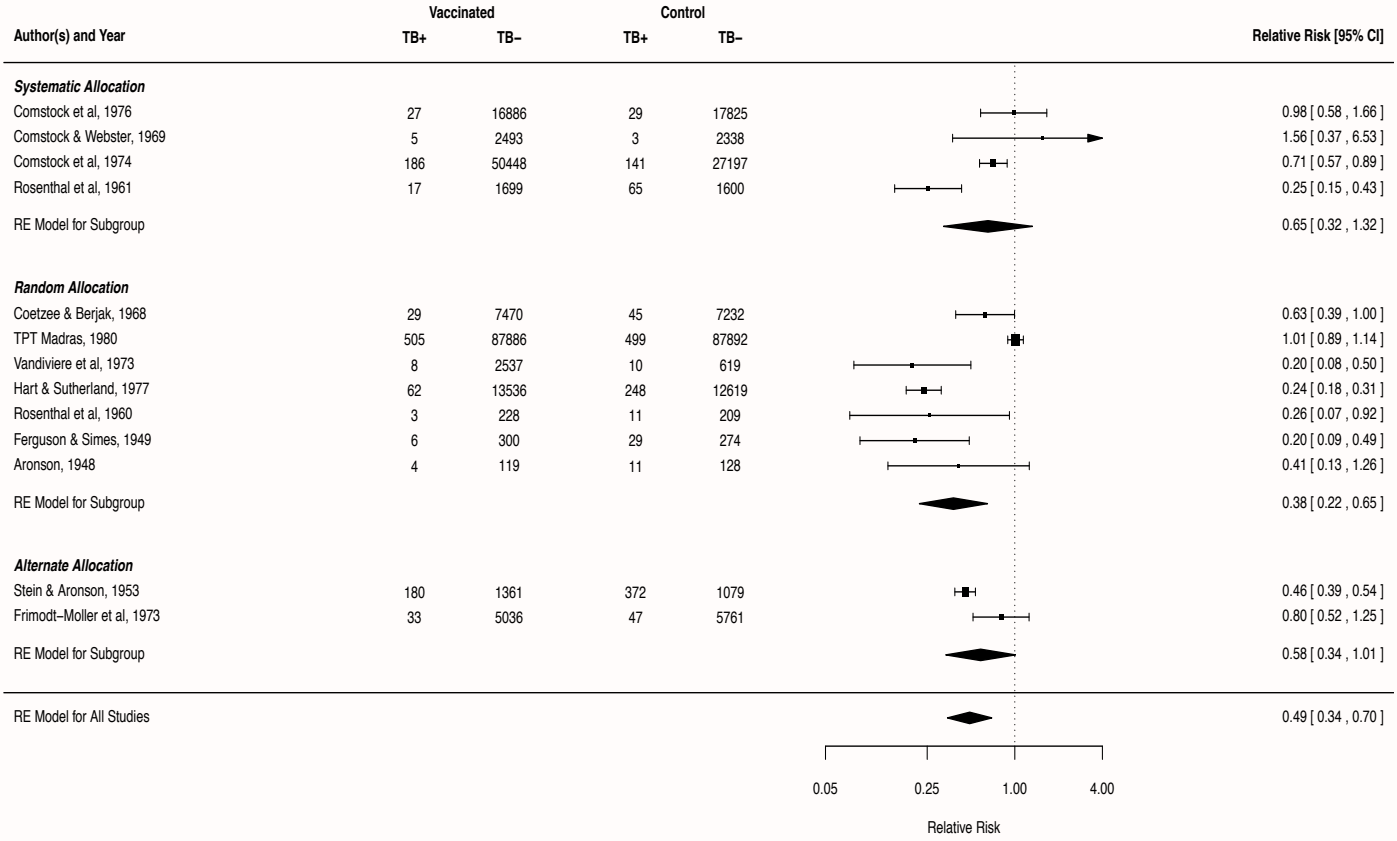
Comparison of Kaplan Meier Survival Curves

R Code

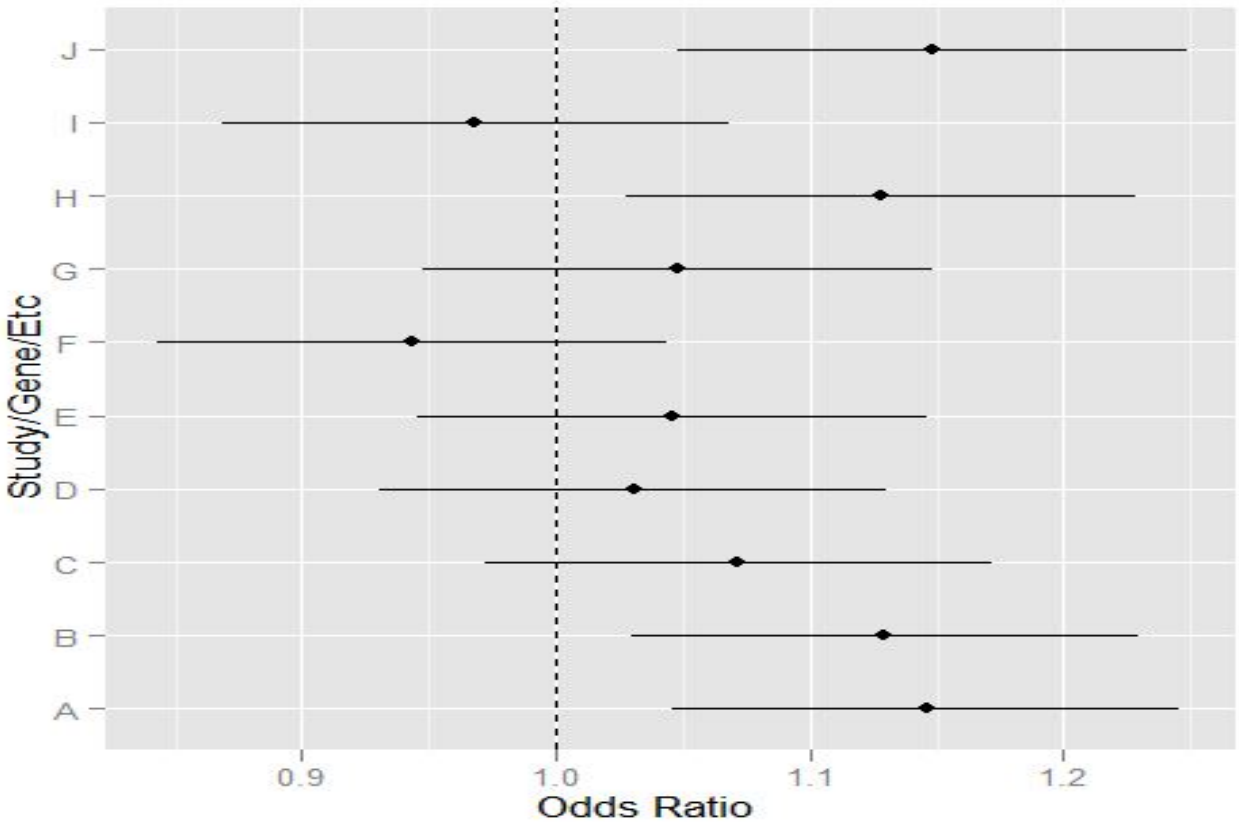
```
> library(ISwR)
> fit.bysex <- survfit(Surv(days, status == 1)~ sex, data=melanom)
> plot(fit.bysex,conf.int=TRUE,col=c("red","blue"),lty=1:2,main ='Comparing Sur ...
> legend(1000,0.2,  c("Male","Female"),lty=c(2,1),lwd=c(1,1),col=c("blue","red"))
```



Forest plot using R package 'metafor'



Forest plot using R package 'ggplot2'



Summary

- R is useful in Data visualization
- Various graphics environment in R
- R can be used to explore data and create various graphics
- Various R functions from various packages for same purpose

Conclusion

From examples we have seen, how graphs generated in R are good – in terms of appearance, time taken for coding and overall quality of the graph. So, R can be an ideal choice for data visualization (exploring) and creating graphs.

Cytel

Thanks for your attention!