Data Visualization and Graphics in R

Xin Luo

December 6th, 2016
Outline

- Plotting using build in graphics tools in R
- Plotting with graphic packages in R (ggplot2)
- Visualizing data by different types of graphs in R (scatter plot, line graph, bar graph, histogram, boxplot, pie chart, heat map, Venn diagram, correlation plot)
- Generate and output polished graphs for publication and presentation
Why using R for plotting

1. When the large sample size exceed the capacity for excel, prism or other graphic tools
2. Fast and simple
3. Super easy with any modification
4. Reproduce the figures and keep exact same format for new figures on new data
5. Bundle with complicated statistical analysis
6. Many add-on packages for various needs beyond build in graphics tools
Basic Plot Example

data()
str(pressure)
plot(pressure)
Labels and Axes

Default: R uses the variable names for axes labels and computes range for axes.

Manual change by:
  • axes labels: xlab, ylab
  • size of labels: cex.lab
  • axes range: xlim, ylim

Titles
  • main: sets plot title (above plot)
  • sub: sets subtitle (beneath plot)

Symbols, colors and lines
  • type: "p" for points, "l" for lines, "b" for both, "h" for histogram-like
  • pch: point symbol
  • col: color
  • cex: size factor
  • lty: line type
  • lwd: line width
Plot symbols
The default color palette in R:

- rainbow(n)
- heat.colors(n)
- terrain.colors(n)
- topo.colors(n)
- cm.colors(n)

R Color Palette other options

RcolorBrewer Package Palette
R color cheat sheet

plot(pressure, col="#0000FF")
plot(pressure, col="blue")
Line Types

Line Types: Ity=

1 2 3 4 5 6
Plot example 1 plot points with formatting

plot (pressure, type="p")

plot (pressure, type="p", pch = 8, cex =0.8, col="red")
Plot example 2 line graph with formatting

```
plot (pressure, type="l")
```

```
plot (pressure, type="l", lty = 3, lwd = 2, col="blue")
```
Plot example 3 add title and text

```r
plot(pressure, main="Relation")

plot(pressure)
text(150,200, label=" p value = 0.05 ")
```
Pie chart

Pie chart is used to visualize the composition of the data groups

pie(table(iris$Species), col=rainbow(3))
data(iris) # load iris data
pch.vec <- c(2, 8, 21)[iris$Species]
col.vec <- c(2, 3, 6)[iris$Species]
plot(iris$Sepal.Length, iris$Sepal.Width, col = col.vec, pch=pch.vec, xlab="sepal.length", ylab="sepal.width",main="iris")
legend ("topleft", pch=c(2, 8, 21), col=c(2, 3, 6), legend = unique(iris$Species), cex=0.8)
Beyond simple graphs: ggplot2

- Hadley Wickham’s ggplot2 package provides a unified interface and simple set of options.

- Once you learn how ggplot2 works for one type of plot, you can easily apply the knowledge for any other types of plots.

- It provides beautiful, publication ready results.

- Easy to plot for data with multiple groups and build legend automatically.


http://www.cookbook-r.com/Graphs/
Build-in R Plotting VS ggplot2

```r
plot(diamonds$carat, diamonds$price, col = diamonds$color,
     pch = as.numeric(diamonds$cut))
```

```r
ggplot(diamonds, aes(carat, price, col = color, shape = cut)) + geom_point()
```
Scatter plot
display the relationship between two continuous variables

```r
ggplot(iris, aes(x=Sepal.Length, y=Sepal.Width, shape=Species, colour=Species)) + geom_point()
```
Scatter plot
Change the points shape and color

```r
ggplot(iris, aes(x=Sepal.Length, y=Sepal.Width, shape=Species, colour=Species)) + geom_point() +
  scale_colour_brewer(palette="Dark2")+
  scale_shape_manual(values=c(2,8,0)) +
  labs(x="Length", y="Width", title="Sepal Length and Width") +
  theme(plot.title = element_text(hjust = 0.5))
```
Scatter plot

Add regression line

ggplot(iris, aes(x=Sepal.Length, y=Sepal.Width, shape=Species, colour=Species)) +
geom_point() +
geom_smooth(method=lm, se=FALSE)
Line Graph
the trend over time or other continuous variables

```r
ggplot(pressure, aes(x=temperature, y=pressure)) + geom_line() + geom_point()
```
Bar Graph

display numeric values (y-axis) for different categories (x-axis)

1. Bar graph for exact value for y

```r
ggplot(ToothGrowth,aes(x=factor(dose),y=len, fill=supp)) +
geom_bar(stat="identity", position="dodge", width=0.5)
```
Bar Graph

2. bar graph for counts of a categorical variable
Bar Graph
3. bar graph for percentage of a categorical variable

ggplot(dig, aes(x= factor(CVD), group=factor(TRTMT))) +
  geom_bar(aes(y = ..prop.., fill = factor(TRTMT)), stat="count", position="dodge") +
  geom_text(aes(label = scales::percent(..prop..), y= ..prop.. ),
            stat = "count",position=position_dodge(0.9), vjust = -.5) +
  labs(x="CVD", y = "Percent", fill="treatment") +
  scale_y_continuous(labels=scales::percent)
Bar Graph

4. Plot mean and error bars

```r
ggplot(ToothGrowth, aes(factor(dose), len )) +
  stat_summary(fun.y = mean, geom = "bar", width=0.5, fill="lightgreen") +
  stat_summary(fun.data = mean_cl_normal, geom = "errorbar", width=0.2)+
  labs(x="dose", y="length")
```
Visualize the distribution of data: 1. histogram

```r
ggplot(ChickWeight, aes(x=weight, fill=factor(Diet))) + geom_histogram(position="identity")
```
Visualize the distribution of data: II. Density Curve

ggplot(ChickWeight, aes(x=weight, fill=Diet)) + geom_density()
Visualize the distribution of data: III. Boxplot

Transformation among histogram, density curve and boxplot

Visualize the distribution of data: III. Boxplot

```r
ggplot(ChickWeight, aes(x=Diet, y=weight)) + geom_boxplot(fill="lightgreen")
```
Visualize the distribution of data: III. Boxplot

```r
ggplot(ChickWeight, aes(x=Diet, y=weight)) + geom_boxplot(fill="lightgreen", notch=TRUE) + stat_summary(fun.y="mean", geom="point", fill="blue", shape=21, size=3)
```
Visualize the distribution of data: III. Boxplot

ggplot(ChickWeight, aes(x=Diet, y=weight)) +
geom_boxplot(notch=TRUE)+
stat_summary(fun.y="mean",geom="point", fill="blue", shape=21, size=3)
Change the overall appearance of the graph
Set background to be black and white

bp1 <- bp + theme_bw()
bp1
Remove grid lines

bp2<-bp1 + theme(panel.grid.major = element_blank(),panel.grid.minor = element_blank())
bp2
Add Labels

bp3 <- bp2 + labs(x="Length", y="Width", title="Sepal Length and Width")
bp3
Modify title and axis labels

bp4 <- bp3 +
theme(axis.title.x = element_text(colour="red", size=11, face="bold"),
     axis.text.x = element_text(colour="blue"),
     axis.title.y = element_text(colour="red", size=11, face="bold", angle = 90),
     axis.text.y = element_text(colour="blue"),
     plot.title = element_text(colour="red", size=12, face="bold", hjust=0.5))

bp4
Modify legend

Sepal Length and Width

bp4 +
theme(legend.background = element_rect(fill="grey85", colour="red", size=0.2),
legend.title = element_text(colour="blue", face="bold", size=11),
legend.text = element_text(colour="red"),
legend.key = element_rect(colour="blue", size=0.2))
Simplified code (plotting for single group)

<table>
<thead>
<tr>
<th>Default tools in R</th>
<th>ggplot</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>plot (x, y, type=&quot;p&quot;)</code></td>
<td>ggplot(data, aes(x=, y=)) + geom_point()</td>
</tr>
<tr>
<td><code>plot (x, y, type=&quot;l&quot;)</code></td>
<td>ggplot(data, aes(x=, y=)) + geom_line()</td>
</tr>
<tr>
<td><code>barplot(x=table(x))</code></td>
<td>ggplot(data, aes(x=factor())) + geom_bar()</td>
</tr>
<tr>
<td><code>hist(x)</code></td>
<td>ggplot(data, aes(x=)) + geom_histogram()</td>
</tr>
<tr>
<td><code>boxplot(data=data, y~x)</code></td>
<td>ggplot(data, aes(x=factor(), y=)) + geom_boxplot()</td>
</tr>
</tbody>
</table>
Simplified code (plotting for multiple groups)

**Scatter plot**

```r
ggplot(data, aes(x= , y=, shape=factor(group), colour=factor(group))) + geom_point()
```

**Bar graph**

```r
ggplot(data, aes(x=factor(), fill=factor(group))) + geom_bar(position=“dodge”)
```

**Histogram**

```r
ggplot(data, aes(x=,fill=factor(group))) + geom_histogram(position =“identity”)
```

**Density Curve**

```r
ggplot(data, aes(x=,fill=factor(group))) + geom_density()
```

**Boxplot**

```r
ggplot(data, aes(x=factor(),y=, fill=factor(group))) + geom_boxplot()
```
# scale data to mean=0, sd=1 and convert to matrix
mtscaled <- as.matrix(scale(mtcars))
# create heatmap and don't reorder columns
heatmap(mtscaled, Colv=F, scale='none')

Heatmap

Heat maps are used to visualize the level of signal of one variable across different time point or other groups
Heatmap

Using ggplot2 to draw heat map on time series data

```
 ggplot(pres_rating, aes(x=year, y=quarter, fill=rating))+geom_tile()
```
Venn Diagram

Venn diagrams are used to check the overlap among different groups

library(Vennerable)
V <- Venn(SetNames=c('A','B','C'),Weight=c(0,10,30,5,20,2,16,1))
plot(V, doWeights=TRUE,type='circles')
Correlation Plot

Correlation plot is used to check the association or similarity among variables.

mcor<-cor(mtcars)
library(corrplot)
corrplot(mcor, method="shade", shade.col=NA, tl.col="black", tl.srt=45)
Output for publication or presentation

Output to PNG:

```r
png("myplot-%d.png", width=400, height=400)
plot(mtcars$wt, mtcars$mpg)
ggplot(mtcars, aes(x=wt, y=mpg)) + geom_point()
dev.off()
```

Output to TIFF:

```r
tiff("myplot-%d.tif", width=400, height=400)
plot(mtcars$wt, mtcars$mpg)
ggplot(mtcars, aes(x=wt, y=mpg)) + geom_point()
dev.off()
```

Output to PDF:

```r
df("myplot.pdf", width=4, height=4)
plot(mtcars$wt, mtcars$mpg)
ggplot(mtcars, aes(x=wt, y=mpg)) + geom_point()
dev.off()
```

Output to postscript:

```r
postscript("myplot.eps", width=4, height=4)
plot(mtcars$wt, mtcars$mpg)
ggplot(mtcars, aes(x=wt, y=mpg)) + geom_point()
dev.off()
```

Note: `dev .off ()` is to let R know you’re finished with plotting commands and it can output the file.
Practice Session

1. An R script that includes all the code covered in this presentation is provided with detailed step-wise annotation and you will go over all the plots during the first half of the practice session.

2. You will try to answer a list of five questions and draw plots to visualize the practicing clinical trial DIG NHLBT Teaching data set.
Q1: Check the relationship between BMI and Systolic BP (SYSBP)

plot(dig$BMI, dig$SYSBP, xlab="BMI", ylab="SYSBP", col="blue")

ggplot(dig, aes(x=BMI, y=SYSBP)) +
  geom_point(colour="blue", shape=21)
Q2: Plot the number of patients for different SEX groups

\[
\text{ggplot}(\text{dig}, \text{aes(x=factor(SEX))) + geom_bar( colour="black", fill="lightgreen", width=0.5)}
\]

\[
\text{barplot(table(dig$SEX), col="lightgreen")}
\]
Q3: Use ggplot to check the distribution of AGE in different treatment groups (TRTMT) using three different types of plots - histogram

ggplot(dig, aes(x=AGE, fill=factor(TRTMT))) + geom_histogram(position="identity")
Q3: use `ggplot` to check the distribution of AGE in different treatment groups (TRTMT) using three different types of plots – density curve.

```r
ggplot(dig, aes(x=AGE, fill=factor(TRTMT))) + geom_density()
```
Q3: use ggplot to check the distribution of AGE in different treatment groups (TRTMT) using three different types of plots – boxplot

ggplot(dig, aes(x=factor(TRTMT), y=AGE)) + geom_boxplot(notch=TRUE, width=0.5, colour="black", fill="lightgreen") + stat_summary(fun.y="mean",geom="point", fill="white", shape=21, size=3)

ggplot(dig, aes(x=factor(TRTMT), y=AGE))  + geom_violin() + geom_boxplot(notch=TRUE, width=0.2, colour="black", fill="lightgreen")+
stat_summary(fun.y="mean",geom="point", fill="white", shape=21, size=3)
Q4. A. plot the percentage of DEATH in different treatment groups (TRTMT)

```r
ggplot(dig, aes(x = factor(DEATH), group=factor(TRTMT))) +
  geom_bar(aes(y = ..prop.., fill = factor(TRTMT)), stat="count", position="dodge")+
  geom_text(aes(label = scales::percent(..prop..), y = ..prop..), stat = "count", position = position_dodge(0.9), vjust = -.5) +
  labs(x = "DEATH", y = "Percent", fill = "treatment") +
  scale_y_continuous(labels = scales::percent)
```
Q4. B. plot the percentage of death attributed to worsening heart failure (DWHF) in different treatment groups (TRTMT)

```r
ggplot(dig, aes(x = factor(DWHF), group=factor(TRTMT))) +
  geom_bar(aes(y = ..prop.., fill = factor(TRTMT)), stat="count", position="dodge") +
  geom_text(aes(label = scales::percent(..prop..), y= ..prop.. ), stat = "count", position = position_dodge(0.9), vjust = -.5) +
  labs(x="DWHF", y = "Percent", fill="treatment") +
  scale_y_continuous(labels = scales::percent)
```
Q5 take plot from Q4b, try to polish the graph

Q4B <- ggplot(dig, aes(x = factor(DWHF), group = factor(TRTMT))) +
  geom_bar(aes(y = ..prop.., fill = factor(TRTMT)), stat = "count", position = "dodge") +
  geom_text(aes(label = scales::percent(..prop..), y = ..prop..), stat =
    "count", position = position_dodge(0.9), vjust = -.5) +
  labs(x = "DWHF", y = "Percent", fill = "treatment") +
  scale_y_continuous(labels = scales::percent)
Q4B +
  theme_bw() +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank()) +
  labs(x = "DWHF", y = "Percentage", title = "Percentage of DWHF in Different Treatment Group") +
  theme(axis.title.x = element_text(colour = "red", size = 11, face = "bold"),
    axis.text.x = element_text(colour = "blue"),
    axis.title.y = element_text(colour = "red", size = 11, face = "bold", angle = 90),
    axis.text.y = element_text(colour = "blue"),
    plot.title = element_text(colour = "red", size = 12, face = "bold", hjust = 0.5)) +
  theme(legend.background = element_rect(fill = "grey85", colour = "red", size = 0.2),
    legend.title = element_text(colour = "blue", face = "bold", size = 11),
    legend.text = element_text(colour = "red"))