R programming for beginners

Ni Shuai

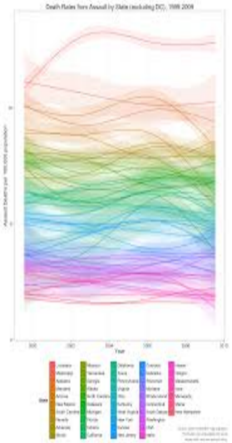
Computational Genome Biology
German Cancer Research Center (DKFZ)

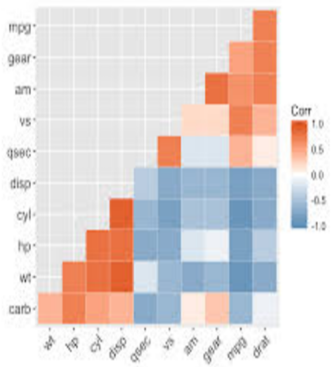
November, 2016

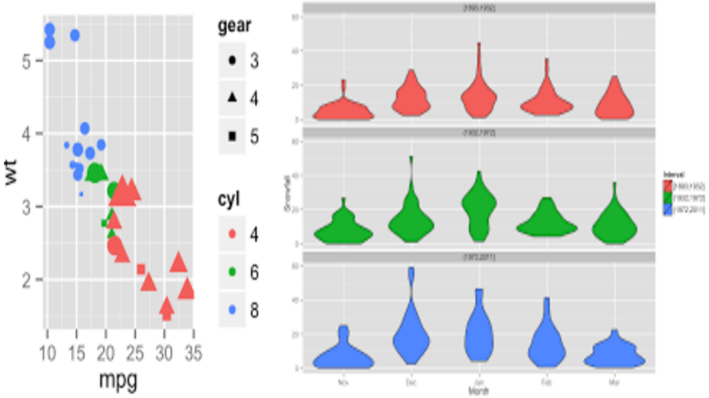
ggplot

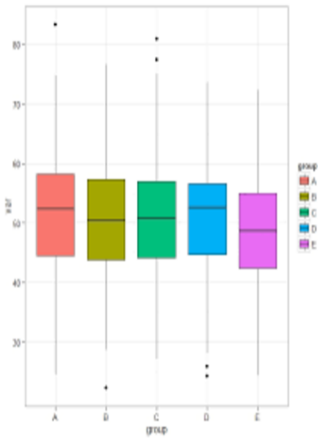
ggplot - a new plotting system in R

- Intrinsically nice looking
- Powerful and smart
- Complete plot system and consistent gramma
- Complicated with simple plot, simple with complicated plot
- Activly matained and developed









ggplot

Essencial components in a ggplot

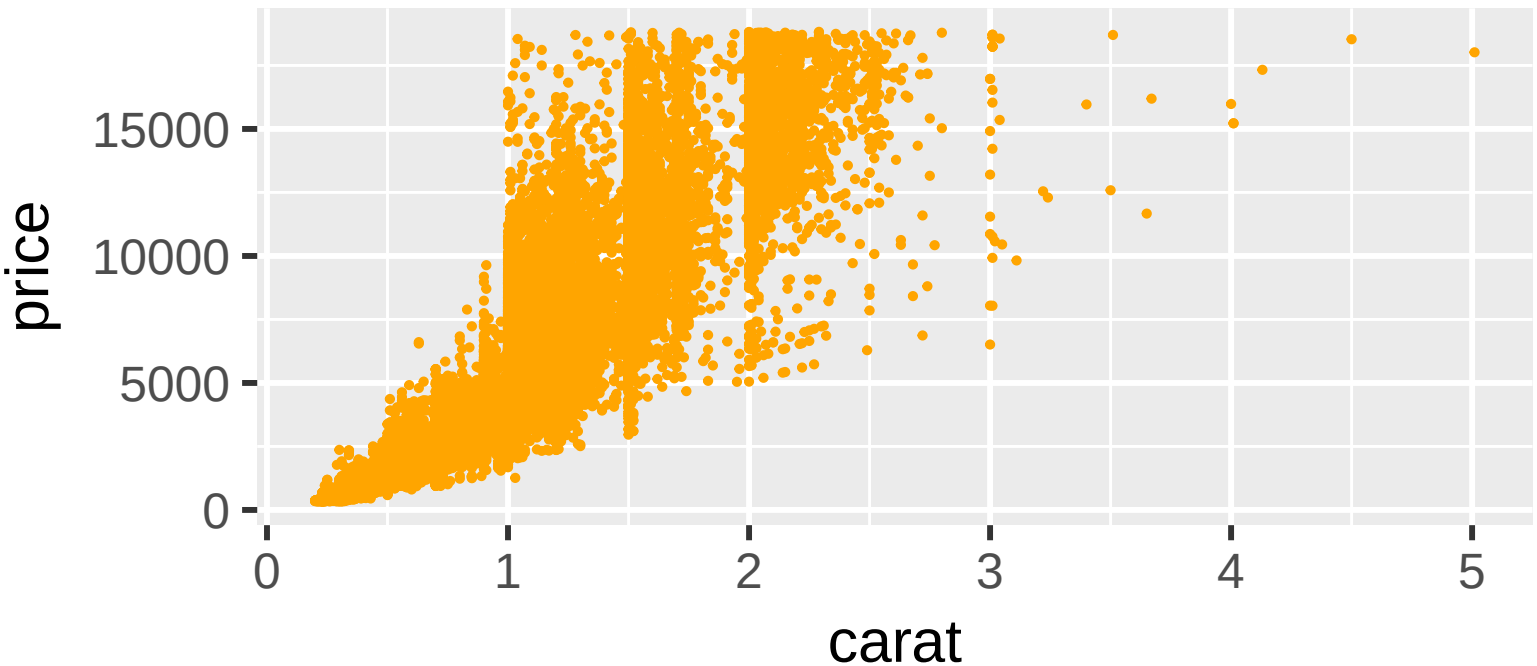
- Input dataset, should always be a data.frame
- X and Y axis - mapping, grouping, coloring
- Layer: the geometric object (plot type)
- Layer: statistical representation of the data
- Position adjustment: dodge, jitter, stack
- Annotataions: addons, lines, borders
- Scales: axis, limits, colors
- Themes: existing themes

The layer geom and stat is always exchangable in different situations based on the plot emphasis

Gramma

All ggplot2 plots start with a call to ggplot(), supplying the input dataset, specifiying aesthethic mappings by aes(). Then add layers, scales, coords and facets with + to the ggplot(). To save a plot to disk, use ggsave().

```
library(ggplot2)
ggplot(data=diamonds, aes(x=carat, y=price)) +
    geom_point(color='orange', size= 0.2)
## ggsave('Myggplot.pdf')
## ggsave("Myggplot.png")
```



Aesthetic mapping

Aesthetic mappings is the central part of the plot, it describes what variable in the data to be represented and should the plot elements be grouped by variables.

```
##city miles per gallon and highway miles per gallon
ggplot(data=mpg, aes(x=cty, hwy))
p = ggplot(data=mpg, aes(x=cty, hwy))
summary(p)

p + geom_point()
```

Aesthetic mappings can also be set in every geom() layers, it may inherit or override the aesthetic mapping from ggplot().

```
ggplot(mpg)+ geom_point(aes(x=cty, y=hwy))
ggplot(mpg, aes(x=cty, y=hwy))
```

Aesthetic mapping

aes() is also used to set color and size by variables in dataset

```
ggplot(data=mpg, aes(x=cty, hwy))+
     geom_point(aes(color=cty))

ggplot(data=mpg, aes(x=cty, hwy))+
     geom_point(aes(color=factor(cyl), size=cyl))+
    ggtitle('title of my first graph')
```

You can also map aesthetics to functions of variables

```
ggplot(data=mpg, aes( x = cty^ 2, y = hwy / cyl))+
    geom_point()
```

Aesthetic mapping

There can be only one variable in aes(), with suitable geom method

```
ggplot(mtcars, aes(mpg))
ggplot(mtcars, aes(mpg))+geom_histogram(binwidth=5)

ggplot(mtcars, aes(mpg))+geom_point()
```

geom lables will override color sceme from the main ggplot call

```
ggplot(mtcars, aes(x=wt, y=mpg, color=cyl))+
    geom_point(size=5, color='green')
ggplot(mtcars, aes(x=wt, y=mpg, color=factor(cyl)))+
    geom_point(size=5)
```

Layers

We can set another layer of statistical representation, variable names should always be inside aes()

```
ggplot(mpg, aes(x=cty, y=hwy))+
   geom_point(aes(color=factor(year), size=displ))+
   stat_smooth()
```

```
ggplot(mpg, aes(x=cty, y=hwy))+
   geom_point(aes(color=factor(year), size=displ),
              alpha=0.5, position= 'jitter')+
   stat_smooth()+
   scale_color_manual(values=c('gold','lightblue'))
```

Aesthetic mapping will not be shared between additional layers

```
ggplot(mpg)+ geom_point(aes(x=cty, y=hwy))+stat_smooth()
```

Layers

Adjust the position of overlaping geom objects

```
##
g= ggplot(diamonds, aes(cut))
g+ geom_bar(aes(fill=color))
g+ geom_bar(aes(fill=color), position='dodge')
g+ geom_bar(aes(fill=color), position='jitter')
```

Scale controls the detail of the data to be prensented, like colors, lables, x and y axis limits and color schemes

```
g= ggplot(mtcars, aes(mpg,disp))
g+ geom_point() +
    labs(title = "Relationship between displacement and miles per gallon",
    y = "Engine displacement")
g+ geom_point() +
    ylim(0, 600)
g+ geom_point(aes(alpha=mpg))
```

Themes

Themes control the display of all non-data elements of the plot. One can use existing themes, or choose to tweak individual settings by using theme()

```
g+ geom_point() +
    theme_dark()

g+ geom_point() +
    theme_classic()
Mytheme=theme(panel.grid.major = element_blank(),
              panel.grid.minor = element_line(color="dodgerblue"),
              panel.background = element_rect(fill='pink'),
              axis.line = element_line(colour = "red"),
              axis.text = element_text(angle = 90, hjust = 1))
g+ geom_point() +
    Mytheme
```

Plot Example

- Bar chart

```
# Number of diamonds in each clarity class:
g <- ggplot(diamonds, aes(clarity))
g + geom_bar()
g + geom_bar(aes(fill=cut))
g + geom_bar(aes(fill=color))+ coord_flip()
```

- Box plot

```
# Distribution of diamonds carat:
g <- ggplot(diamonds, aes(x=clarity, y=carat))
g + geom_boxplot()
g + geom_boxplot(aes(fill=clarity))
g + geom_boxplot(aes(fill=color))
```

Plot Example

- Deisity plot

```
# Density plot for diamond price
g=ggplot(data = diamonds, aes(x =price))
g+geom_density()
g+geom_density(aes(color=color))
```

- Histgram

```
# Histgram for diamond carat
ggplot(diamonds, aes(price)) +
  geom_histogram()
ggplot(diamonds, aes(carat)) +
  geom_histogram(binwidth = 0.01)
ggplot(diamonds, aes(price)) +
  geom_histogram(aes(fill=clarity))
```

Plot Example

- Line chart

```
scale_fill_manual(values=rep(brewer.pal(8, 'Pastel1')[c(1,2,5)], 5))
set.seed(100)
rainfall=data.frame(matrix(rnorm(48), 8, 6))
rainfall=rainfall+5
names(rainfall)=c('Jan','Feb','Mar','Apr','May','Jun')
rainfall[1:4,]=rainfall[1:4,] + rep(seq(0,9,length.out = 8), each=6)
rainfall$city=c('Beijing','Bangkok','Delhi',
                'Moscow','Suzhou','Lima', 'Berlin','Madrid')

rainfall=melt(rainfall)
ggplot(rainfall, aes(variable, value, color=city))+
    geom_line(aes(group=city), size=2 )+
    geom_point(size=3)+
     theme(panel.grid.minor=element_blank(),
          panel.grid.major=element_blank(),
          panel.background=element_blank(),
          panel.border=element_blank())+
    scale_fill_manual(values=brewer.pal(8, 'set2'))
```

Exercises

- Use a boxplot to show the relationships between Miles/(US) gallon (mpg) and Number of cylinders (cyl) of all cars
  - Color the boxes by number of cylinders
- use a violin plot to show the relationships between Miles/(US) gallon (mpg) and Number of cylinders (cyl) of all cars
  - add jittered dots to the violinplot, with color indicating different weights of cars (wt)
  - add jittered dots to the violinplot, with shape indicating different types of transmissions (am)

Exercises

- Study the relationships between caret, price and clarity for a randomly selected 200 records in diamond dataset
  - plot carat vs. price with color represented by diamond clarity(clarity)
    - with a smooth curve with confidence intervals to the plot
    - with a linear model regression line with confidence intervals to the plot
  - explore the distribution of diamond carat using density plot
  - explore the price distribution for each set of clarity with a density plot
  - explore the price distribution for each set of clarity with a jittered dot plot
  - explore the price - clarity distribution for diamonds with price greater than 2000 US dollars