

Getting started with ggplot2

STAT 133

Gaston Sanchez

Department of Statistics, UC–Berkeley

`gastonsanchez.com`

`github.com/gastonstat/stat133`

Course web: `gastonsanchez.com/stat133`

ggplot2

Resources for "ggplot2"

- ▶ Documentation: <http://docs.ggplot2.org/>
- ▶ Book: **ggplot2: Elegant Graphics for Data Analysis** (by Hadley Wickham)
- ▶ Book: **R Graphics Cookbook** (by Winston Chang)
- ▶ RStudio ggplot2 cheat sheet

<https://www.rstudio.com/wp-content/uploads/2015/03/ggplot2-cheatsheet.pdf>

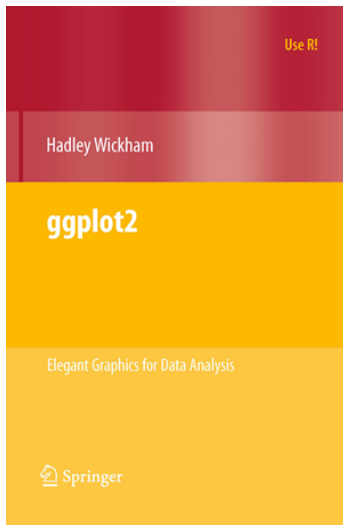
package "ggplot2"

```
# remember to install ggplot2  
# (just once)  
install.packages("ggplot2")
```

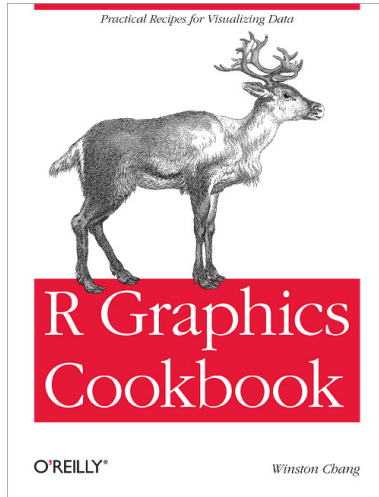
```
# load ggplot2  
library(ggplot2)
```

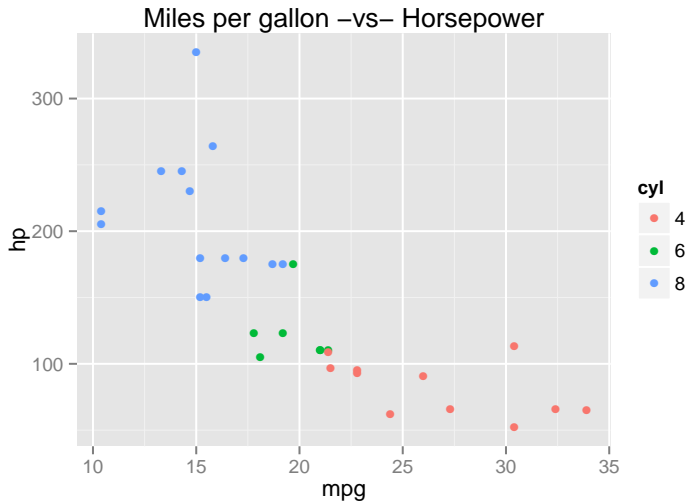
```
# see basic documentation  
?ggplot
```

ggplot2 book

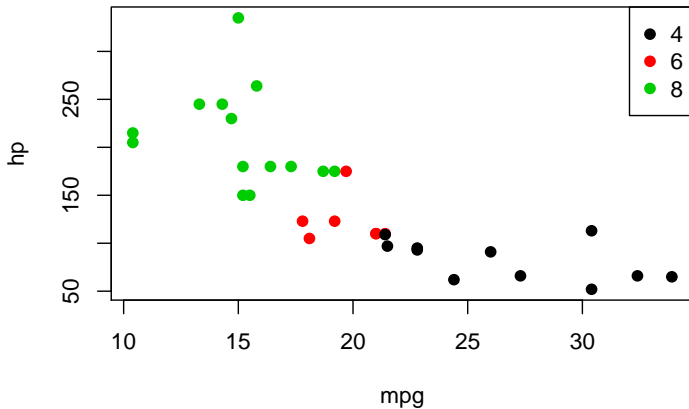


R Graphics Cookbook





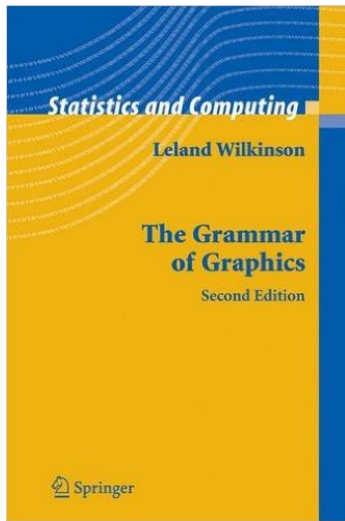
Miles per gallon –vs– Horsepower



About "ggplot2"

- ▶ "ggplot2" (by Hadley Wickham) is an R package for producing statistical graphics
- ▶ It provides a framework based on Leland Wilkinson's **Grammar of Graphics**
- ▶ "ggplot2" provides beautiful plots while taking care of fiddly details like legends, axes, colors, etc.
- ▶ "ggplot2" is built on the R graphics package "grid"
- ▶ Underlying philosophy is to describe a wide range of graphics with a compact syntax and independent components

The Grammar of Graphics



About the Grammar of Graphics

- ▶ *The Grammar of Graphics* is Wilkinson's attempt to define a theoretical framework for graphics
- ▶ **Grammar:** Formal system of rules for generating graphics
 - Some rules are mathematic
 - Some rules are aesthetic

About the Grammar of Graphics

3 Stages of Graphic Creation

- ▶ **Specification:** link data to graphic objects
- ▶ **Assembly:** put everything together
- ▶ **Display:** render of a graphic

About the Grammar of Graphics

Specification

Link data to graphic objects

- ▶ Data
- ▶ Transformation of variables (e.g. aggregation)
- ▶ Scale transformations (e.g. log)
- ▶ Coordinate system (e.g. cartesian)
- ▶ Graphic Elements (e.g. points, lines)
- ▶ Guides (e.g. labels, legends)

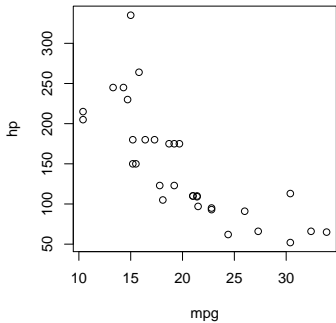
R package "ggplot2"

About "ggplot2"

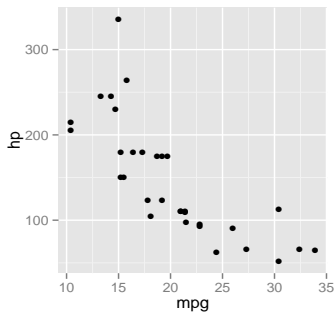
- ▶ Default appearance of plots carefully chosen
- ▶ Designed with visual perception in mind
- ▶ Inclusion of some components, like legends, are automated
- ▶ Great flexibility for annotating, editing, and embedding output

Base graphics -vs- "ggplot2"

base graphics



ggplot2



About "ggplot2"

- ▶ "ggplot2" is the name of the package
- ▶ The gg in "ggplot2" stands for *Grammar of Graphics*
- ▶ Inspired in the **Grammar of Graphics** by Lee Wilkinson
- ▶ "ggplot" is the class of objects (plots)
- ▶ ggplot() is the main function in "ggplot2"

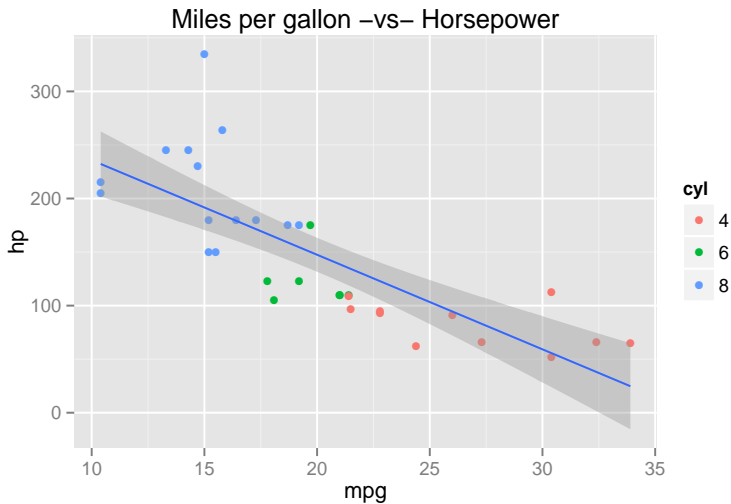
What is a Statistical Graphic?

Some Data set

mtcars

##	mpg	hp	cyl
## Mazda RX4	21.0	110	6
## Mazda RX4 Wag	21.0	110	6
## Datsun 710	22.8	93	4
## Hornet 4 Drive	21.4	110	6
## Hornet Sportabout	18.7	175	8
## Valiant	18.1	105	6
## Duster 360	14.3	245	8
## Merc 240D	24.4	62	4
## Merc 230	22.8	95	4
## Merc 280	19.2	123	6

What is a statistical graphic?



What is a statistical graphic?

Elements to draw the chart “manually”

What is a statistical graphic?

Elements to draw the chart “manually”

- ▶ coordinate system
- ▶ x and y axis (intervals)
- ▶ axis tick marks
- ▶ axis labels, and title
- ▶ points (with colors)
- ▶ regression line (and ribbon)
- ▶ legend

What is a statistical graphic?

Simply put, a statistical graphic is:

- ▶ A mapping from data to aesthetic attributes (color, shape, size) of geometric objects (points, lines, bars)
- ▶ A plot may also contain statistical transformations of the data
- ▶ A plot is drawn on a specific coordinate system
- ▶ Sometimes faceting can be used to get the same plot for different subsets of the dataset

Starting with "ggplot2"

starwarstoy.csv

```
## Warning in file(file, "rt"): cannot open file  
'/Users/gaston/Documents/stat133/stat133/datasets/starwarstoy.csv': No  
such file or directory  
## Error in file(file, "rt"): cannot open the connection  
## Error in eval(expr, envir, enclos): object 'starwars' not found
```


Scatterplot

```
## Error in ggplot(data = starwars):  object 'starwars' not found
```

Main steps in creating ggplot graphics

1 Dataset

A	B	C	D	E	F

2 Which variables

A	B	C	D	E	F

3 Geometric objects

● *points*

abcd *text*

~ *lines*

■ *bars*

4 Aesthetics

x = A

y = B

color = C

size = *default*

shape = *default*

Building a scatterplot

User specifications

- ▶ Dataset: `starwars`
- ▶ Variables: `height`, `weight`, `jedi`
- ▶ Geoms: `points`
- ▶ Aesthetics (attributes):
 - `x`: `height`
 - `y`: `weight`
 - **`color`**: `jedi`

Scatterplot with "ggplot2"

```
ggplot(data = starwars) +  
  geom_point(aes(x = height, y = weight, color = jedi))
```

Scatterplot with "ggplot2"

```
ggplot(data = starwars) +  
  geom_point(aes(x = height, y = weight, color = jedi))
```

- ▶ `ggplot()` initializes a "ggplot" object
- ▶ specify the dataset with `data`
- ▶ type of geometric object: `geom_point()`
- ▶ mapping aesthetic attributes to variables with `aes()`
 - x-position: `height`
 - y-position: `weight`
 - color: `jedi`

Scatterplot with "ggplot2"

```
ggplot(data = starwars) +  
  geom_point(aes(x = height, y = weight, color = jedi))
```

```
## Error in ggplot(data = starwars):  object 'starwars' not found
```

Scatterplot with "ggplot2"

Automated things in "ggplot2"

- ▶ Axis labels
- ▶ Legends (position, labels, symbols)
- ▶ Choose of colors for points
- ▶ Background color (e.g. gray)
- ▶ Grid lines (major and minor)
- ▶ Axis tick marks

you can always change the automated elements

"ggplot2" graphics

Philosophy of "ggplot2"

A graphic is a **mapping** from **data** to **aesthetic attributes** (color, shape, size) of **geometric objects** (points, lines, bars)

Scatterplot with "ggplot2"

```
ggplot(data = starwars) +  
  geom_point(aes(x = height, y = weight, color = jedi))
```

```
## Error in ggplot(data = starwars): object 'starwars' not found
```

Mapping

data values

height	weight	jedi
1.72	77	jedi
1.50	49	no_jedi
1.82	77	jedi
1.80	80	no_jedi
0.96	32	no_jedi
1.67	75	no_jedi
0.66	17	jedi
2.28	112	no_jedi



aesthetic attributes

x	y	color
x_1	y_1	#F8766D
x_2	y_2	#00BFC4
x_3	y_3	#F8766D
x_4	y_4	#00BFC4
x_5	y_5	#00BFC4
x_6	y_6	#00BFC4
x_7	y_7	#F8766D
x_8	y_8	#00BFC4

"ggplot2" graphics

Philosophy of "ggplot2"

A graphic is a **mapping** from **data** to **aesthetic attributes** (color, shape, size) of **geometric objects** (points, lines, bars)

- ▶ `ggplot(data, ...)`
- ▶ `aes()`
- ▶ `geom_objects()`

Scatterplot with "ggplot2"

How does "ggplot2" work?

- ▶ plots are created piece-by-piece
- ▶ plot components added with **+** operator
- ▶ aesthetic attributes mapped to data values
- ▶ computation of scales for aesthetic attributes

How does it work?

Usually, we specify the data and variables inside the function `ggplot()`

```
ggplot(data = mtcars, aes(x = mpg, y = hp))
```

Note the use of the internal function `aes()` to *map* x to mpg, and y to hp.

Then we **add a layer** of geometric objects: points in this case

```
+ geom_point()
```

Some alternative options

```
# option A  
ggplot(data = starwars,  
       aes(x = height, y = weight, color = jedi)) +  
  geom_point()
```

Some alternative options

```
# option A  
ggplot(data = starwars,  
       aes(x = height, y = weight, color = jedi)) +  
  geom_point()
```

```
# option B  
ggplot(data = starwars) +  
  geom_point(aes(x = height, y = weight, color = jedi))
```

Some alternative options

option A

```
ggplot(data = starwars,  
       aes(x = height, y = weight, color = jedi)) +  
  geom_point()
```

option B

```
ggplot(data = starwars) +  
  geom_point(aes(x = height, y = weight, color = jedi))
```

option C

```
ggplot() +  
  geom_point(data = starwars,  
            aes(x = height, y = weight, color = jedi))
```


Main inquiries

Always ask yourself ...

- ▶ What is the **data** set of interest?
- ▶ What **variables** will be used to make the plot?
- ▶ What **graphics shapes** will be used to display?
- ▶ What **features** of the shapes will be used to represent the data values?

"ggplot2" basics

- ▶ The data must be in a `data.frame`
- ▶ Variables are mapped to aesthetic attributes
- ▶ Aesthetic attributes belong to geometric objects **geoms** (points, lines, polygons)

Basic Terminology

- ▶ **ggplot()** - The main function where you specify the dataset and variables to plot
- ▶ **geoms** - geometric objects
 - `geom_point()`, `geom_bar()`, `geom_line()`, `geom_density()`
- ▶ **aes** - aesthetics (i.e. attributes)
 - shape, color, fill, linetype

Warning

"ggplot2" comes with the function `qplot()` (i.e. *quick plot*).
Avoid using it!

As Karthik Ram says: “you’ll end up unlearning and relearning a good bit”