Package ggplot2

1 Presentation du package

ggplot2 est un systeme de creation declarative de graphiques, base sur The Grammar of Graphics. Vous fournissez les donnees, dites a ggplot2 comment mapper les variables a l'esthetique, quelles primitives graphiques utiliser et il s'occupe des details.

2 Installation du package ggplot2

Comme tout package, il faut commencer par l'importer. Pour cela il y a plusieurs solutions. Personnellement, j'utilise l'onglet package sous R Studio (dans la fenetre en bas a droite), puis le sous onglet install

Une fois que le package est installe, il faut le charger :

library(ggplot2)

3 Principes de fonctionnement du package ggplot2

Le package ggplot2 fonctionne par couche successives. La premiere d'entre elles, est un peu le canevas du graph. Elle consiste a indiquer, dans quel jeu de donnees se trouve les donnees, et quelles sont les variables que l'on souhaite representer. Ensuite, une seconde couche est ajoutee, elle consiste, par exemple, a indiquer le type de graph que l'on souhaite realiser : scatterplot, boxplot, barplot etc. Viennent ensuite les couches d'affinage en quelque sorte, qui vont permettre de choisir les couleurs, les echelles des axes, les options de legende etc.

3.1 Definition de la couche canevas

Pour definir ce que j'appelle la couche canevas, on utilise la fonction "ggplot()" et son argument "aes().". Les graphs construits avec ggplot2 commencent toujours par ce type de ligne de code :

ggplot(database, aes(x=, y=))

3.2 Definition du type de plot : geom_XXX

Il s'agit ensuite de definir le type de graph que l'on souhaite realiser : un scatter plot, un boxplot, un barplot, etc. Pour cela, on rajoute un signe plus en bout de la premiere ligne (celle du canevas), et on ajoute une nouvelle ligne avec la fonction adequate : geom_point() pour un scatter plot, geom_boxplot() pour un boxplot, geom_bar() pour un barplot etc.

Et pour connaitre toutes les fonctions geom_XXX disponibles, elles sont decrites dans la partie "Geoms" de la cheatsheet du package ggplot. Vous pouvez la telecharger automatiquement en allant dans l'onglet Help -> Cheatsheets -> Data Visualization with ggplot2.

3.3 Definition des options du graph

Dans un troisieme temps, on affine le graph, en precisant differentes couches concernant :

les echelles des axes : avec la fonctions scale_x_continuous() les couleurs : avec la fonction scale_colour_manual() les noms des axes : avec les fonctions xlab(), ylab() la legende avec la fonction theme(legend.position,=“bottom”)

4.Realiser un scatterplot avec ggplot2

4.1 Le scatterplot de base

Imaginons que l’on souhaite realiser un scatter plot avec le jeu de donnees iris, en representant la base des donnees iris la variable Sepal.Length en y et la variable Sepal.Width en x :

ggplot(database, aes(x=Sepal.Width, y=Sepal.Length))+ geom_point()

4.2 Definir des couleurs selon une variable

Le jeu de donnees comporte, en realite, trois especes d’iris differentes (variable Species). Pour representer les points avec une couleur differente par espece, on va definir l’argument colour dans la fonction aes() de la partie “canevas” :

ggplot(database, aes(x=Sepal.Width, y=Sepal.Length, colour=Species))+ geom_point()

4.3 Utiliser des formes de points differentes

Pour cela, on utilise l’argument shape dans aes() : ggplot(database, aes(x=Sepal.Width, y=Sepal.Length, colour=Species, shape=Species))+ geom_point()+ scale_colour_manual(values=c(“magenta”, “orange”, “blue”))

5 Realiser un boxplot avec ggplot 2

5.1. Boxplot de base

Imaginons que l’on souhaite realiser un boxplot de la variable Sepal.Length, par espece. Pour cela, on utilise la fonction geom_boxplot().

ggplot(database, aes(y=Sepal.Length,x=Species))+ geom_boxplot()

6 Realiser des barplot avec ggplot2

Il existe deux types de barplot realisables avec ggplot 2. Les premiers, que j’appelle “barplot de comptage”, permettent de representer un nombre de donnees dans chaque modalite d’une variable.

La seconde categorie consiste a representer un parametre statistique comme une moyenne. Personnellement je suis completement opposee ce type de graph, car ils ne permettent pas de visualiser le nombre de donnees, ni leur repartition, ni de la presence eventuelle d’outliers ! Dans cette situation il est preferable de faire un boxplot (surtout qu’il est possible d’y faire figurer la moyenne en plus de la mediane).

Dans tous les cas, on utilise la fonction geom_bar().

6.1 Barplot de comptage

En utilisant le jeu de donnees “mtcars”, imaginons, par exemple, que je veuille representer le nombre de voitures ayant 3,4 ou 5 vitesses (variables gear) :

pas de y car c'est un comptage

ggplot(database,aes(as.factor(gear)))+ geom_bar()

J'utilise ici factor(gear) car cette variable est consideree comme une variable numerique

6.2 Barplot en representant un parametre statistique

Par exemple, ici je vais representer les moyennes de la variable mpg (miles per gallon) pour tous les croisements des modalites des variables gear (nombre de vitesses) et carb (nombre de carburateurs). Pour cela, il est necessaire de fournir a la fonction ggplot, un jeu de donnees comportant ces moyennes. Une facon tres simple de les obtenir est d'utiliser la fonction summarySE() du package Rmisc.

library(Rmisc)

summarySE(database,measurevar="mpg", groupvars=c("gear","carb")

7. Le facetting

C'est une des grandes possibilites de ggplot2. Ca consiste a sous diviser un graph, selon les modalites d'une ou plusieurs variables. Ici par exemple, je vais utiliser le jeu de donnees Melanoma du package MASS, et je vais etudier les relations entre la variable thickness et time, pour chacune des modalites de la variable status :

library(MASS)

ggplot(database, aes(y=thickness, x=time))+ geom_point()+ geom_smooth()+ facet_wrap(~status)