الكلية متعددة التخصصات تارودانت
ⵜⴰⴽⵍⵉⵜ ⵜⴰⵖⵜⵉⵙⵉⵢⵜ ⵏ ⵜⵓⵣⵍⴰⵏⵜ
FACULTE POLYDISCIPLINAIRE TAROUDANT

Le 22/10/2022

Master : Big Data et Intelligence Artificielle

Semestre 3

Big data 2

ggplot2 package

Presented by Naaima BEN KADOUR
Supervised by :
Professor Mohamed El Hajji
Professor Tariq AIT BAHA

Academic Year 2022-2023

What is ggplot2?

ggplot2 is an open-source data visualization package for the statistical programming language R. it's based on The Grammar of Graphics.

What is the Grammar of Graphics?

A grammar of graphics is basically a tool that enables us to describe the components of a given graphic. Basically, this allows us to see beyond the named graphics, (scatter plot, to name one) and to basically see the underlying statistics behind it. Consider grammar of graphics as the grammar of English where we use different words, tenses, punctuations to form a sentence.

Components of Grammar of graphics

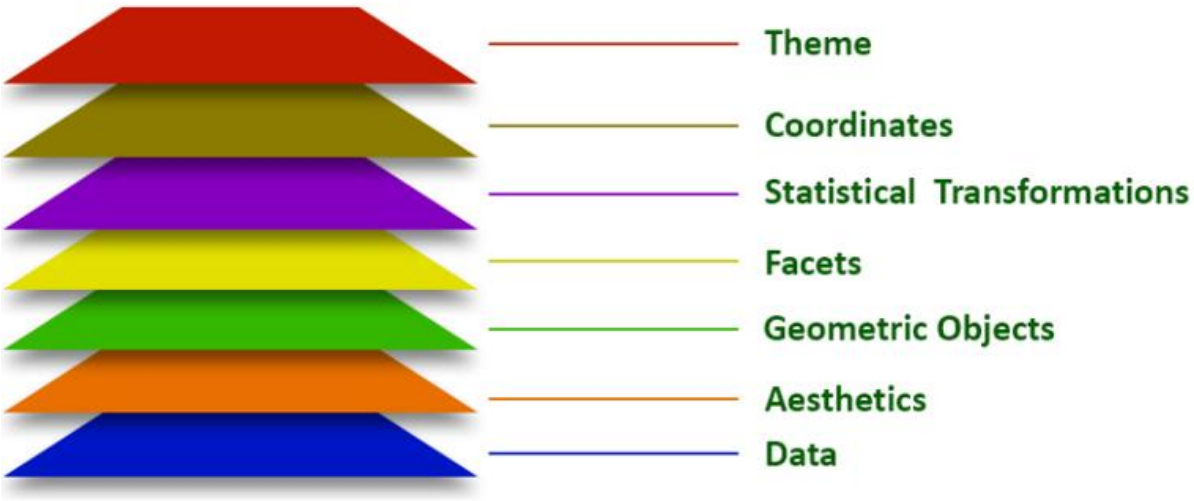
Typically, to build or describe any visualization with one or more dimensions, we can use the components shown in the below image.

the three main components that are required to create a plot, and without these components, the plotline would not be able to plot the graph. These are:

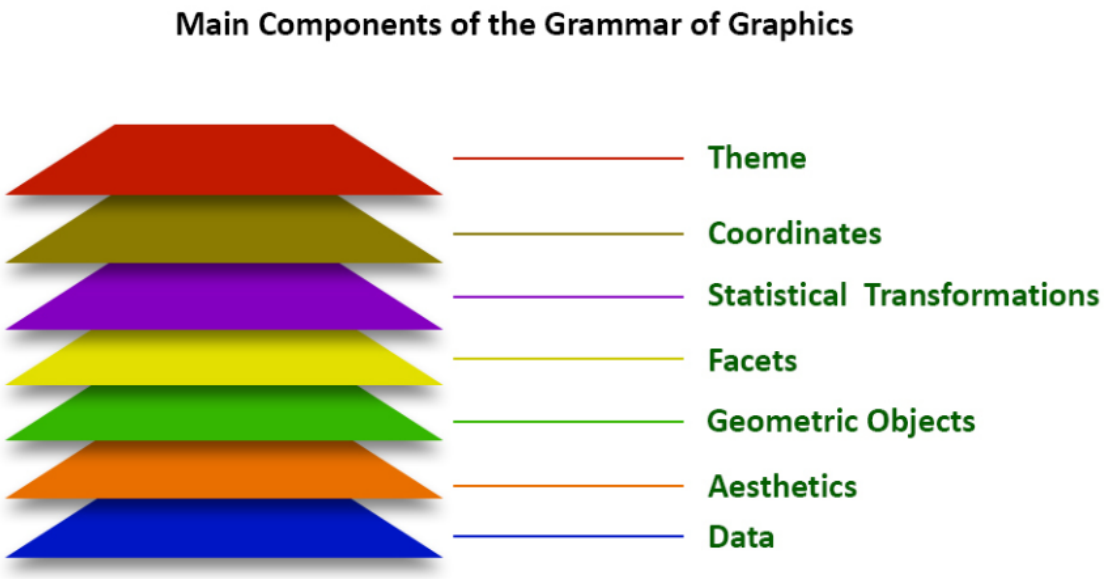
- **Data** is the dataset that is used for plotting the plot.
- **Aesthetics** is the mapping between the data variables and the variables used by the plot such as x-axis, y-axis, color, fill, size, labels, alpha, shape, line width, line type.
- **Geometric Objects** is the type of plot or a geometric object that we want to use such as point, line, histogram, bar, boxplot, etc.

There are various optional components that can make the plot more meaningful and presentable. These are:

- **Facets** allow the data to be divided into groups and each group is plotted separately.
- **Statistical transformations** compute the data before plotting it.



- **Themes** define the presentation of the data such as font, color, etc.



Who created ggplot2?

Hadley Wickham created ggplot2 in 2005, he is a statistician from New Zealand and Chief Scientist at RStudio Inc, an adjunct Professor of statistics at the University of Auckland, Stanford University, and Rice University. He is best known for his development of open-source software for the R statistical programming language for data visualisation, including ggplot2.

When was ggplot2 created?

The initial release of ggplot2 was on 10 June 2007 (15 years ago).

Why is ggplot2 good for data visualization?

ggplot2 is declaratively and efficiently to create data visualization based on **The Grammar of Graphics.** The layered grammar makes developing charts structural and effusive. Generating ggplot2 feels like playing with LEGO blocks. The core concept and syntax are elegant to onboard new users, and the community supports advanced use cases. Various beautiful themes and colour palettes in ggplot2 make the visualization look professional and engaging with the end-users.

How does ggplot2 works?

Example of data visualization using R

Data set: mtcars(motor trend car road test) comprise fuel consumption and 10 aspects of automobile design and performance for 32 automobiles and come pre-installed with dplyr package in R.

```r
# Installing the package
install.packages("dplyr")

# Loading package
library(dplyr)

# Summary of dataset in package
summary(mtcars)
```

Output:

```
      mpg             cyl             disp             hp       
 Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0  
 1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5  
 Median :19.20   Median :6.000   Median :196.3   Median :123.0  
 Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7  
 3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0  
 Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0  
      drat             wt             qsec             vs        
 Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000  
 1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000  
 Median :3.695   Median :3.325   Median :17.71   Median :0.0000  
 Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375  
 3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000  
 Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000  
       am              gear            carb      
 Min.   :0.0000   Min.   :3.000   Min.   :1.000  
 1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000  
 Median :0.0000   Median :4.000   Median :2.000  
 Mean   :0.4062   Mean   :3.688   Mean   :2.812  
 3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000  
 Max.   :1.0000   Max.   :5.000   Max.   :8.000  
```

We devise visualizations on **mtcars** dataset which includes 32 car brands and 11 attributes using **ggplot2** layers.

Data Layer:

In the data Layer we define the source of the information to be visualize, let's use the mtcars dataset in the ggplot2 package

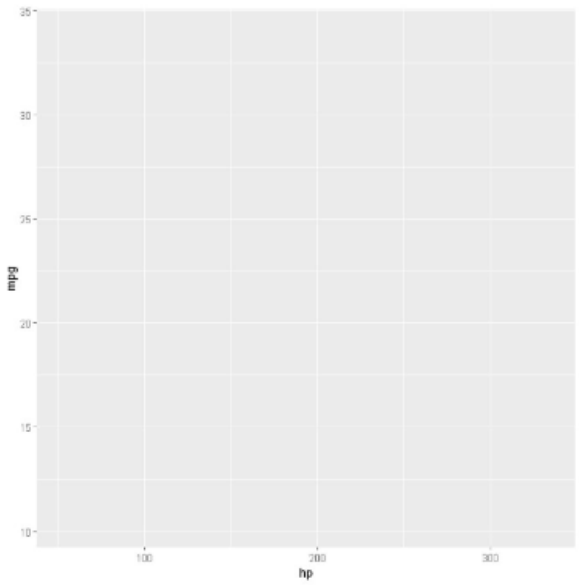
```r
# Loading packages
library(ggplot2)
library(dplyr)

# Data Layer
ggplot(data = mtcars)
```

Aesthetic Layer:

Here we will display and map dataset into certain aesthetics.

```r
# Aesthetic Layer
ggplot(data = mtcars, aes(x = hp, y = mpg, col = disp))
```

Output:



Geometric layer:

In geometric layer control the essential elements, see how our data being displayed using point, line, histogram, bar, boxplot

```
# Geometric layer
ggplot(data = mtcars,
       aes(x = hp, y = mpg, col = disp)) + geom_point()
```

Output: