Introdução a Manipulação e Visualização de dados usando o pacote **tidyverse** R

Prof. Dr. Allan Robert

Minicurso do Encontro Bahiano de Estatística da UFBA.
Universidade Federal da Bahia

19, 21 e 23 de Outubro, 2020

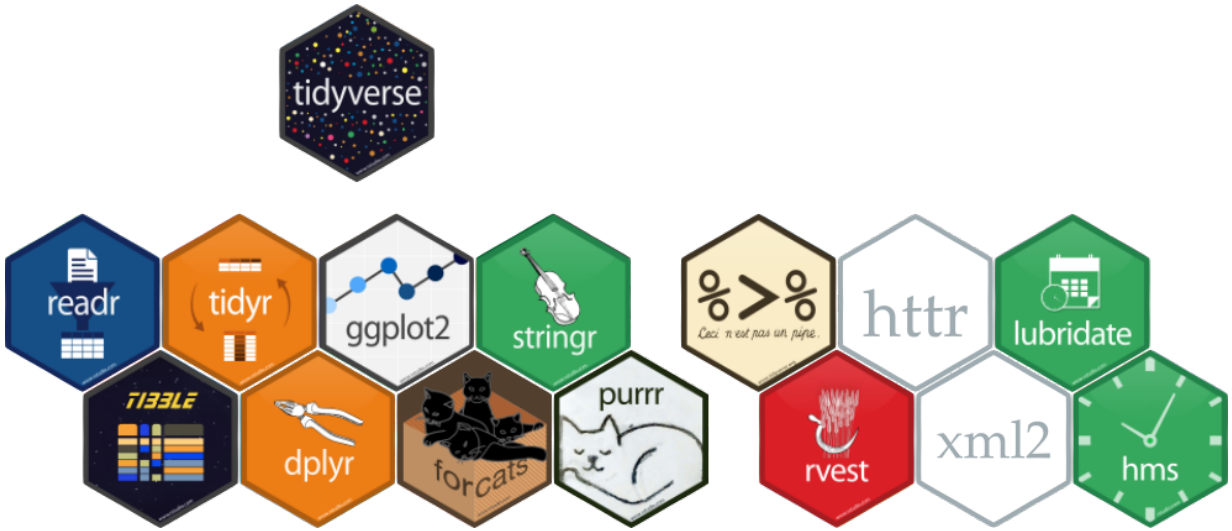
Objetivo

Os objetivos deste minicurso:

- Mostrar a forma de uso pensada no pacote **tidyverse** para manipular e visualizar dados;
- Aplicar esta metodologia usando dados reais e disponíveis publicamente.
- Pensar como resolver problemas reais a partir de uma análise crítica de dados.
- Diminuir o tempo e custo do processo inicial de dados que geralmente é nomeado de Análise (Ou Estatística) Descritiva dos dados.

tidyverse e seus Pacotes

O **tidyverse** na verdade é um conjunto de 8 pacotes que foram criados simultaneamente e que também usa funções de outros pacotes do R no qual se comunica muito bem.



Operador pipe ou pipelines

- Original do pacote **magrittr** %>%.
- Alternativa prática para funções aninhadas que confundem devido ao uso, por vezes, excessivos de parenteses.
- Idéia seja x um vetor e y uma função o comando: x %>% y implica $f_y(x)$.
- analogamente x %>% y %>% k implica $f_k(f_y(x))$.

Operador pipe e exemplos

```
x=rnorm(10,20,4) # Gerando 10 números da distribuição N(20,4)
### aplicando soma e depois a raiz e depois aplicando logaritmo natural
log(sqrt(sum(x))) ### Tradicional

## [1] 2.670894

x %>% sum %>% sqrt %>% log  ### pipe

## [1] 2.670894
```

tibble

- Apresenta a classe tibble para tabelas de dados.
- Melhoria da classe data.frame (padrão do R).
- Apresenta uma visualização bem mais informativo.
- É prático para criação de dados direto no R.

Saida de um data.frame

Exemplo: Dados de plantas já implentados do R
Usando a função head para ver somente as 6 primeiras linhas de dados.
Caso digite direto o nome dos dados o R mostrará todas linhas.

```
head(iris)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

Saida de uma tibble

Transformando em tibble com as_tibble e visualizando.

```
## Transformando em uma tibble e visualizando
iris=as_tibble(iris);iris
```

```
## # A tibble: 150 x 5
##    Sepal.Length Sepal.Width Petal.Length Petal.Width Species
##           <dbl>       <dbl>        <dbl>       <dbl> <fct>
##  1          5.1         3.5          1.4         0.2 setosa
##  2          4.9         3            1.4         0.2 setosa
##  3          4.7         3.2          1.3         0.2 setosa
##  4          4.6         3.1          1.5         0.2 setosa
##  5          5           3.6          1.4         0.2 setosa
##  6          5.4         3.9          1.7         0.4 setosa
##  7          4.6         3.4          1.4         0.3 setosa
##  8          5           3.4          1.5         0.2 setosa
##  9          4.4         2.9          1.4         0.2 setosa
## 10          4.9         3.1          1.5         0.1 setosa
## # ... with 140 more rows
```

readr

Função alternativa para leitura (agrega read.table, read.csv entre outras)

- Focado na importação de dados .txt, .csv e semelhantes.
- Permite fazer a leitura já especificando os tipos de cada variável.
- Na leitura os dados são exportado como uma tibble.
- Reconhece campos com variáveis da classe data.
- Para leituras de outros tipos de dados:
  - **readxl** para planilhas excel;
  - **haven** para SPSS, Stata e SAS.
  - **rvest** para HTML
  - **httr** para Web APIs

tidyr

Foco em funções para organização de dados

- Funções (gather e spreed) para empilhar e desempilhar dados.
- Funções unir e separar dados de variáveis (separate e unite).
- Funções específicas para dados faltantes.
- Funções para combinações de fatores (Muito usado em experimentação).

stringr

Usado para manipulação de caracteres (Que pode ser um problema significativo, pois pessoas alimentam banco de dados de formas diferentes).

- Funções para juntar e separar caracteres.
- Funções contagem de caracteres totais e por condições.
- Funções alterar caracteres sob condições.
- Funções Extrair caracteres de acordo com a posição.
- Entre outras.

forcats

Usado para manipulação de fatores, que podem ser bastante proveitosas para visualização gráfica.

- Ordenar fatores considerando o interesse do estudo.
- Ordenar fatores considerando a frequência desta variável.
- Ordenar fatores considerando a ordem de uma outra variável.
- Ordenar fatores considerando a frequencia de outra variável.

dplyr

Um dos pacotes mais usados e focado na exploração dos dados.

- Funções para juntar (concatenar) tabelas de dados usando chaveamento.
- Agrupar informações por uma categorias de uma variável.
- Manipulação de variaveis no sentido de selecionar, filtrar, renomeaar, etc.
- Extração de medidas de resumo.
- Ainda aceita diversas funções padrões como argumento.

purrr

Focado em programação funcional. A ideia é ter uma função que chama outras funções e para aplicadar aos dados. É uma espécie de aperfeiçoamento da família apply da base do R.

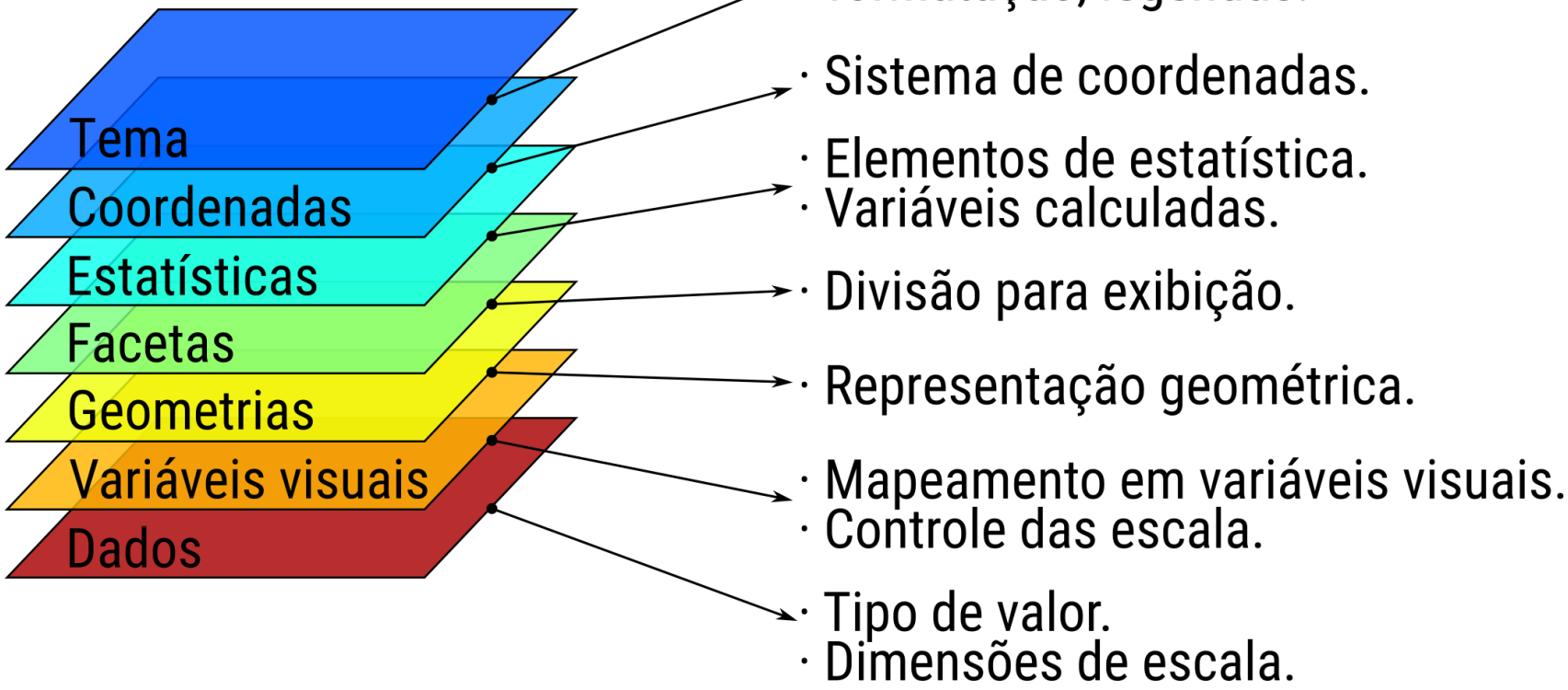
- Funções para aplicar funções da base do R e do tidyverse.
- Resumir informações por categorias de uma variável.
- Aplicar funções para diferentes classes do R.
- Permite formas mais resumidas de programação.
- Aninhamento de dados, no qual pode ser obtidas medidas de resumo entre outras.

ggplot2

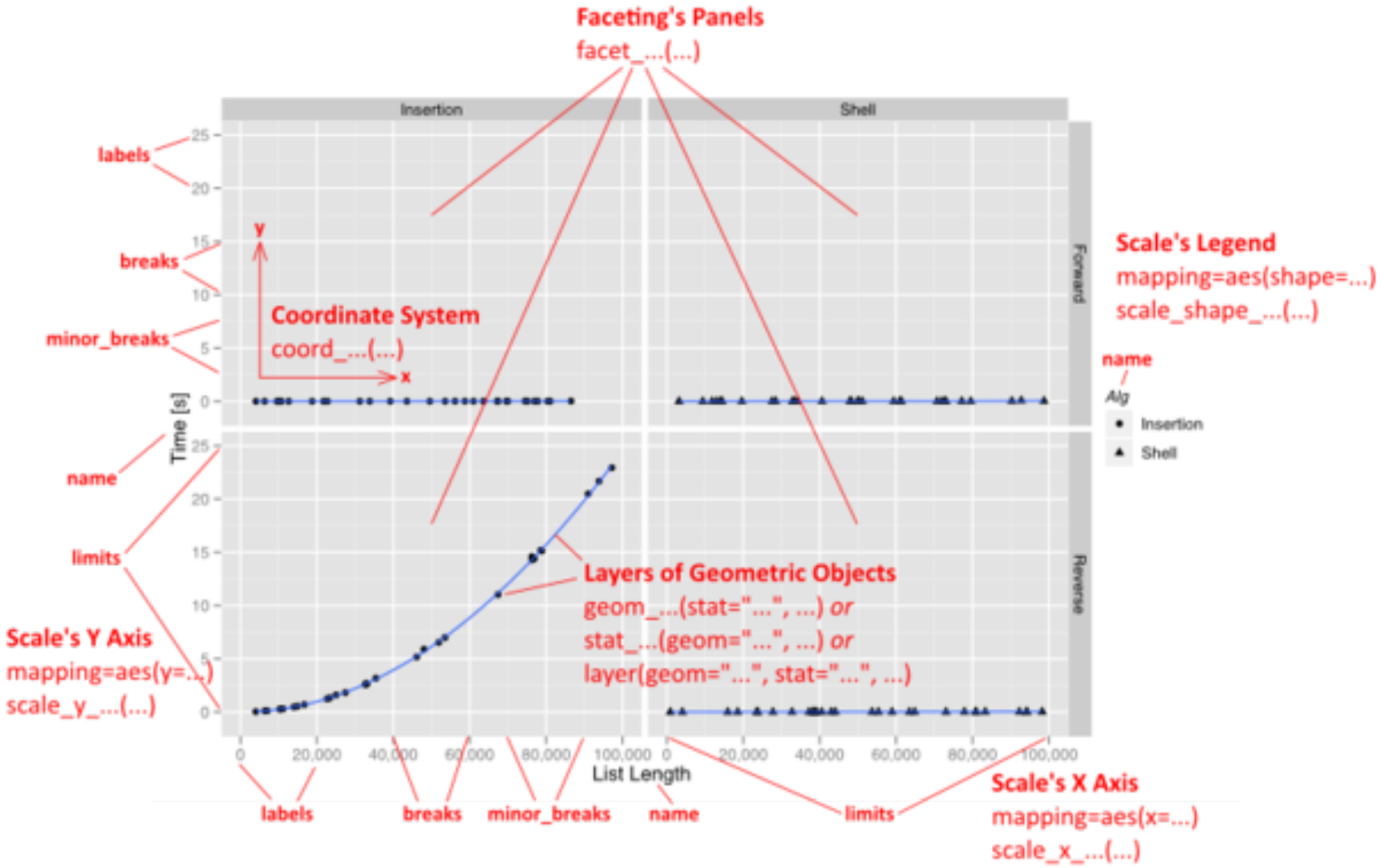
Focado em visualização gráfica. Provavelmente o mais usado de todos os pacotes.

- Apresenta uma visualiza gráfica moderna e elegante.
- Inspirado pelo livro de Wilkinson (2005) intitulado "The Grammar of Graphics" que propõe uma gramática que pode ser usada para descrever e construir uma ampla gama de gráficos a partir de múltiplas camadas de informações.
- Versão apresenta diversos formatos (chamaremos de geoms) para criação de gráficos, podendo ser salvos variando tamanho, extensão e qualidade.
- Atualmente diversos novos pacotes (extensões do ggplot2) foram criados com novos formatos e mesma filosofia (Oficialmente 79).

Descrevendo as camadas

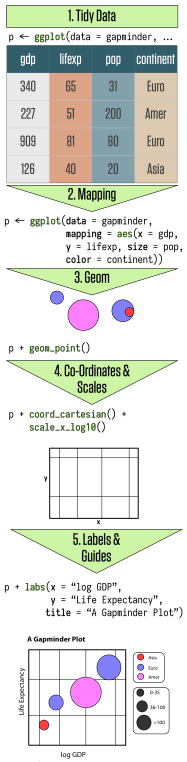


Anatomia gráfica do ggplot2



ver figura original: http://sape.inf.usi.ch/quick-reference/ggplot2

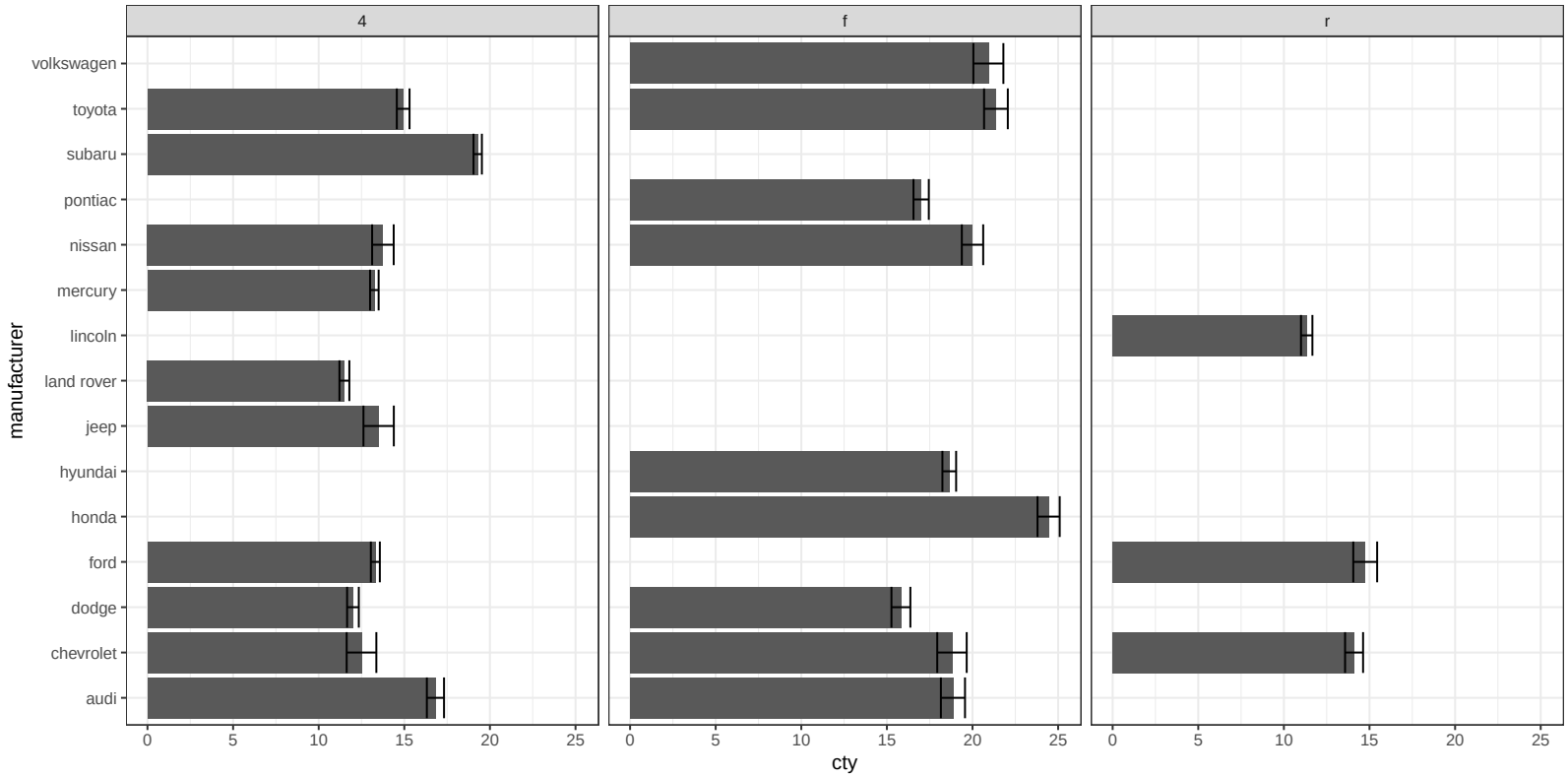
Diagrama das camadas com comandos



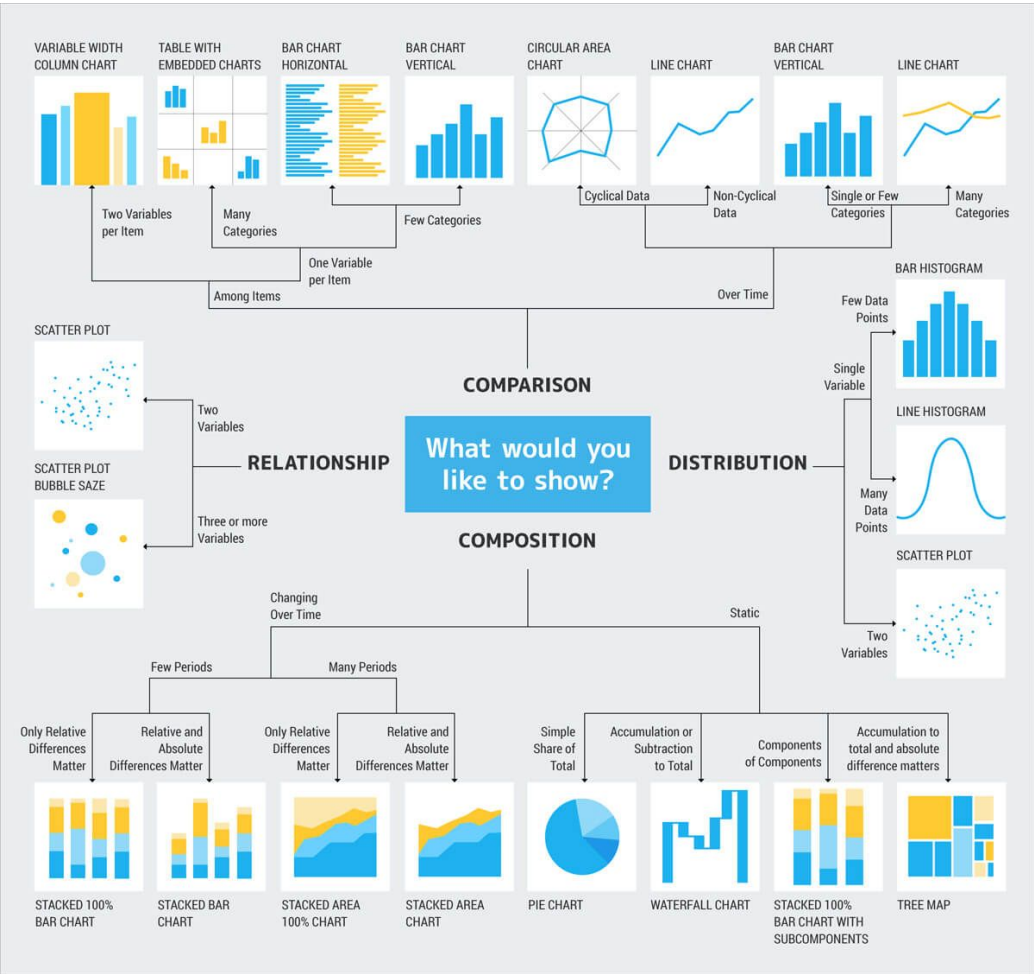
Ver figura original: https://socviz.co/makeplot.html

Exemplo

```
ggplot(mpg, aes(x=manufacturer, y=cty)) + #  1 e 2 camadas dados
geom_bar(stat="summary", fun.y="mean") + # 3 geom (estat "sum" muda para "mean")
facet_grid(.~drv)+ # 4 camada de faces por tipo de marcha
# acrescentado outra camada estatística com mínimo e máximo
geom_errorbar(stat="summary", fun.data="mean_se") +# 5 camada estatística
coord_flip() +# trocando as coordenadas de x e y
theme_bw() # 7 camada de tema petro e branco
```



Algumas opções gráficas



ver figura original: https://extremepresentation.com/design/7-charts/

Bibliografia Usadas

Bibliografia Usadas:

- Introdução ao R- Landeiros.
- R para cientistas sociais
- R for Data Science (https://r4ds.had.co.nz/).
- Advanced R (https://adv-r.hadley.nz/).
- Hadley Wickham. ggplot2 : Elegant Graphics for Data Analysis. Springer, 2009.
- Wilkinson, Leland. The Grammar of Graphics (2n ded.). Statistics and Computing,New York:Springer, 2005.

Sites

Sites muito interessantes:

- https://cran.r-project.org/doc/contrib/Landeiro-Introducao.pdf
- https://www.tidyverse.org/
- https://ggplot2.tidyverse.org/reference/theme.html
- https://exts.ggplot2.tidyverse.org/gallery/
- Introduction tidyverse - PDF
- http://material.curso-r.com/manip/
- http://leg.ufpr.br/ walmes/cursoR/data-vis/index.html
- https://curso-r.github.io/ragmatic-book/principios.html
- https://rstudio.com/resources/cheatsheets/

Agora vamos a parte prática