

Análise de dados com R

uma visão inicial das atividades de um cientista de dados


Rosangela de Fátima Pereira Marquesone – rpereira@larc.usp.br

Francisco Pereira Junior – fpereira@utfpr.edu.br

Tereza Cristina Melo de Brito Carvalho – terezacarvalho@usp.br

04/10/2018

Tópicos

- 
- Introdução ao R
 - Primeiros passos com R
 - Análise de dados com R

Material para atividade prática

Material utilizado

Acessar o conteúdo do material em:

<https://github.com/jolai-r/minicurso>

Arquivos:

- **JolaiMinicursoRDia1.pdf** – Slides utilizados no primeiro dia do minicurso
- **JolaiMinicursoRDia2.pdf** – Slides utilizados no segundo dia do minicurso
- **lastfm.csv** – Base de dados para a atividade de análise de dados

Sugestões para se tornar um cientista de dados

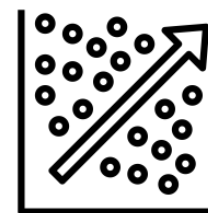
HABILIDADE COM R OU PYTHON



HABILIDADE COM TECNOLOGIAS DE BIG DATA



CONHECIMENTO EM ESTATÍSTICA



Sobre o R

Criado por Ross Ihaka e Robert Gentleman nos anos 90

- Universidade de Auckland, Nova Zelândia



Como o R foi desenvolvido

1. Criado inicialmente para testar ideias estatísticas
2. Utilizado posteriormente como ferramenta de ensino de cursos de estatística
3. Adotado posteriormente por um grande número de usuários e desenvolvedores

Características do R

Divulgado ao público com publicação de um artigo no periódico “*The Journal of Computational Statistics and Graphics*”, em 1996

R: A Language for Data Analysis and Graphics

ROSS IHAKA and Robert GENTLEMAN

In this article we discuss our experience designing and implementing a statistical computing language. In developing this new language, we sought to combine what we felt were useful features from two existing computer languages. We feel that the new language provides advantages in the areas of portability, computational efficiency, memory management, and scoping.

Key Words: Computer language; Statistical computing.

Características do R

R oferece suporte à diversas operações de análise de dados

- Classificação
- Agrupamento
- Análise de série temporal
- K-Means
- Análise exploratória
- Mineração de texto
- Modelos lineares e não lineares
- **Muito mais...**

Características do R

Exemplos de soluções disponíveis no CRAN - The Comprehensive R Archive Network - <https://cran.r-project.org/>

Nome	Descrição
arm	Pacote para criação de modelos lineares
igraph	Pacote para análise de redes. Utilizado para representar redes sociais
lubridate	Pacote com diversas soluções para manipulação de datas
reshape	Pacote para agregação de dados
tm	Pacote para mineração de texto. Adequado para atuar com dados não estruturados
XML	Pacote para manipulação de arquivos XML e HTML

13.121 pacotes disponíveis

Características do R

R é aplicado em diferentes áreas

- Finanças
- Ciências sociais
- Genética
- Medicina
- Redes sociais
- Marketing
- Varejo
- Telecomunicações
- ...

Características do R

R é utilizado por diversas empresas



The New York Times

Características do R

Benefícios do R

- Plataforma **única** para análise de dados

Ambiente R

Manipulação
de dados

Análise de
dados

Visualização
de dados

Características do R

Benefícios do R

- Permite integração com diversos bancos de dados



Arquivos de
texto



Banco de dados
relacionais



Banco de dados
distribuídos

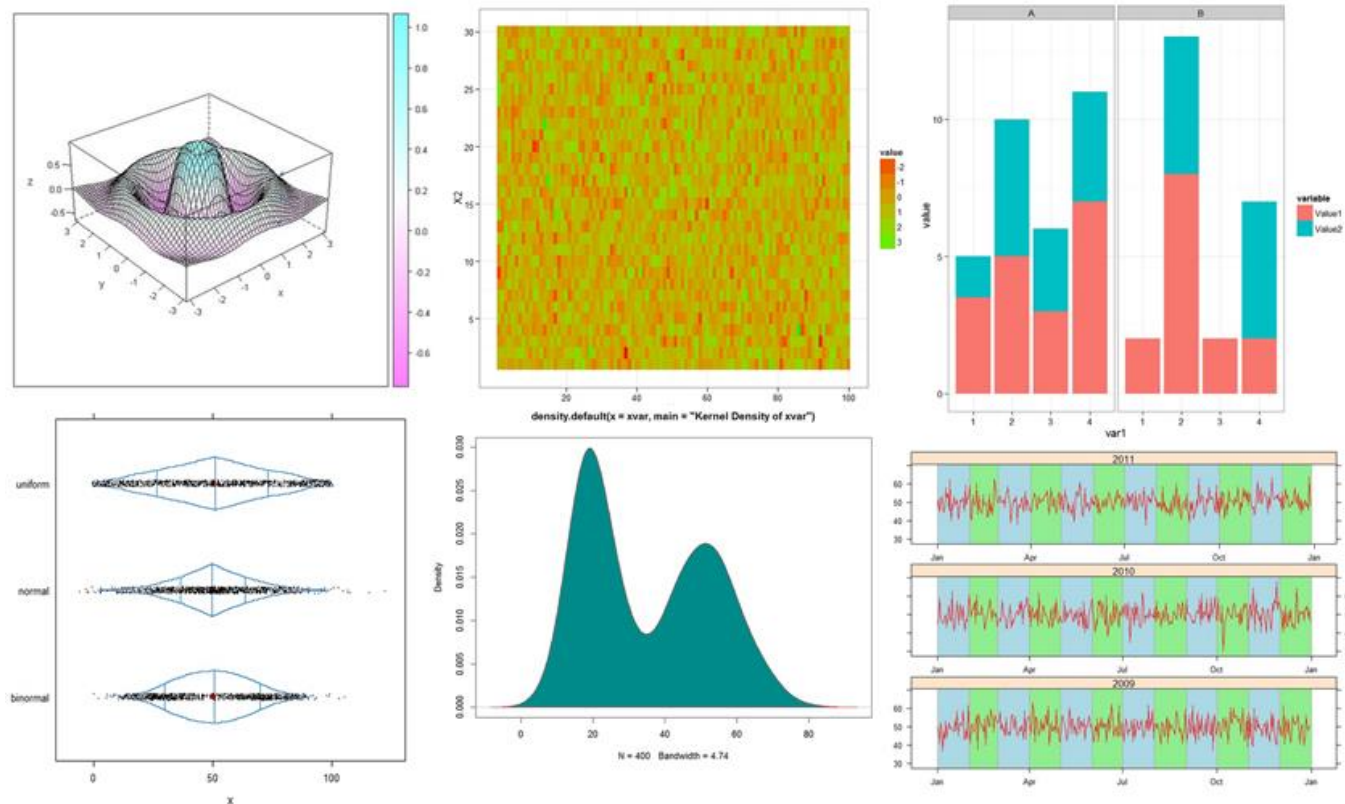


Streaming de
dados

Características do R

Benefícios do R

- Grande variedade de recursos para visualização de dados



Tópicos



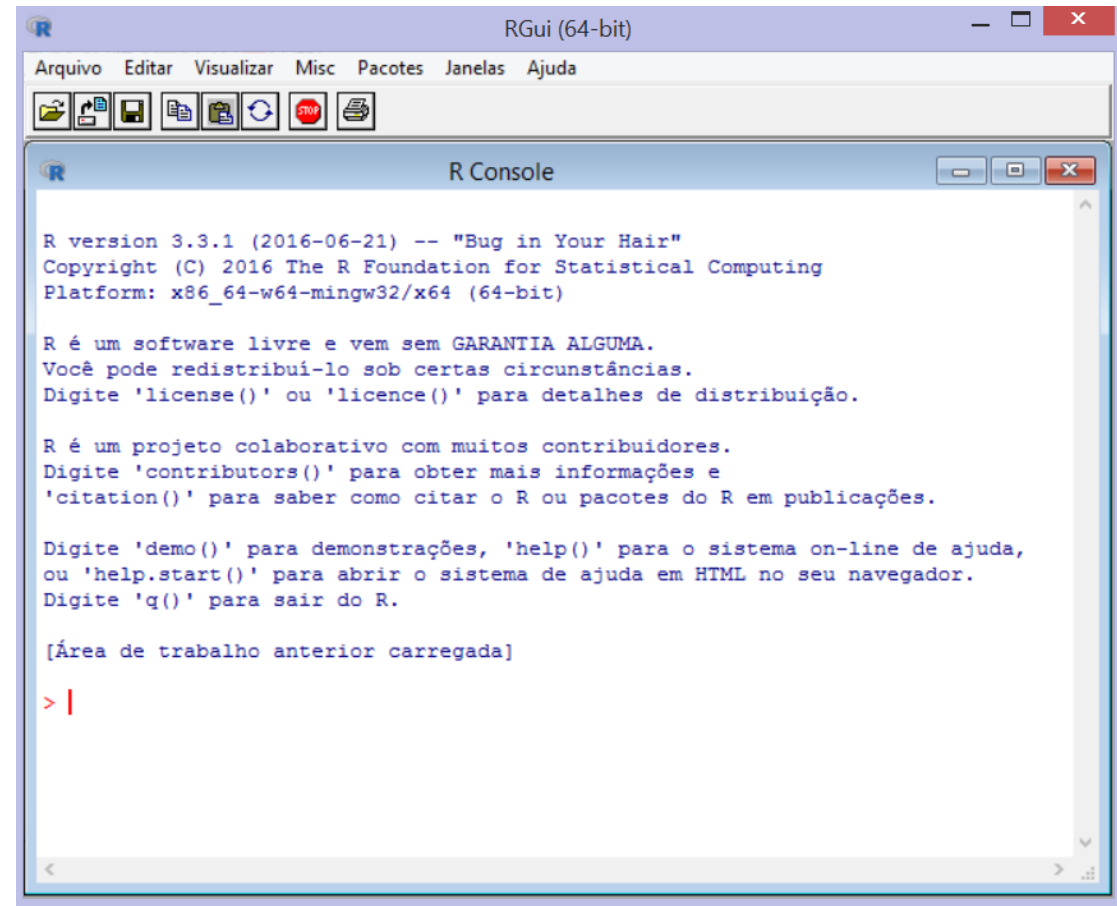
Introdução ao R

Primeiros passos com R

Análise de dados com R

Primeiros passos com R

Ferramenta utilizada

A screenshot of the RGui (64-bit) window. The window has a menu bar with 'Arquivo', 'Editar', 'Visualizar', 'Misc', 'Pacotes', 'Janelas', and 'Ajuda'. Below the menu bar is a toolbar with icons for file operations. The main area is the 'R Console', which displays the following text:

```
R version 3.3.1 (2016-06-21) -- "Bug in Your Hair"
Copyright (C) 2016 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R é um software livre e vem sem GARANTIA ALGUMA.
Você pode redistribuí-lo sob certas circunstâncias.
Digite 'license()' ou 'licence()' para detalhes de distribuição.

R é um projeto colaborativo com muitos contribuidores.
Digite 'contributors()' para obter mais informações e
'citation()' para saber como citar o R ou pacotes do R em publicações.

Digite 'demo()' para demonstrações, 'help()' para o sistema on-line de ajuda,
ou 'help.start()' para abrir o sistema de ajuda em HTML no seu navegador.
Digite 'q()' para sair do R.

[Área de trabalho anterior carregada]

> |
```


Primeiros passos com R

- R permite a execução de operações aritméticas
- Exemplos:

```
> 5 + 100
```

```
[1] 105
```

```
> 200 - 80
```

```
[1] 120
```

```
> 10 * 10
```

```
[1] 100
```

```
> 50 / 5
```

```
[1] 10
```

```
> 2^4
```

```
[1] 16
```

```
> 5 %% 2
```

```
[1] 1
```

Primeiros passos com R

- R permite a utilização de operadores lógicos
- Exemplos:

```
> 20 > 10
```

```
[1] TRUE
```

```
> 20 >= 20
```

```
[1] TRUE
```

```
> 30 > 50
```

```
[1] FALSE
```

```
> 10 >= 9
```

```
[1] TRUE
```

```
> 8 == 5
```

```
[1] FALSE
```

```
> 20 != 30
```

```
[1] TRUE
```

Primeiros passos com R

- O símbolo de atribuição é o “<-”

- Exemplo:

```
> varX <- 4
```

```
> varY <- varX + 4
```

- A visualização do conteúdo de uma variável pode ser visualizada digitando seu nome:

```
> varY
```

```
[1] 8
```

Primeiros passos com R

- R permite a chamada de funções
- Exemplo:

```
> print("Eu amo R")  
[1] "Eu amo R"
```

```
> class(varY)  
[1] "numeric"
```

```
> sqrt(4)  
[1] 2
```

Primeiros passos com R

- R permite a construção de funções definidas pelo usuário
- Exemplo:

#construindo a função

```
> minhafuncao <- function(x, y){  
  soma <- x + y  
  return(soma)  
}
```

#fazendo a chamada da função

```
> minhafuncao(10, 20)  
[1] 30
```

Primeiros passos com R

- R permite o uso de “data frame”, uma estrutura de dados para armazenar tabelas de dados.
- Exemplos:

#visualizando a estrutura de um dataframe com registros de veículos

> mtcars

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.070	17.40	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.730	17.60	0	0	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.780	18.00	0	0	3	3

Primeiros passos com R

- Exemplos de operações em um data frame:

#buscando o registro do data frame na posição de linha 1 e coluna 2

```
> mtcars[1,2]
```

```
[1] 6
```

#verificando o número de linhas

```
> nrow(mtcars)
```

```
[1] 32
```

#acessando o conteúdo de uma coluna do dataframe

```
> mtcars$gear
```

```
[1] 4 4 4 3 3 3 3 4 4 4 4 3 3 3 3 3 4 4 4 3 3 3 3 3 4 5 5 5 5 5 4
```

Primeiros passos com R

- Exemplos de operações em um data frame:

#verificando a estrutura do data frame

> str(mtcars)

```
'data.frame': 32 obs. of 11 variables:
 $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
 $ cyl : num   6  6  4  6  8  6  8  4  4  6 ...
 $ disp: num  160 160 108 258 360 ...
 $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
 $ drat: num   3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
 $ wt  : num   2.62 2.88 2.32 3.21 3.44 ...
 $ qsec: num   16.5 17 18.6 19.4 17 ...
 $ vs  : num   0  0  1  1  0  1  0  1  1  1 ...
 $ am  : num   1  1  1  0  0  0  0  0  0  0 ...
 $ gear: num   4  4  4  3  3  3  3  4  4  4 ...
 $ carb: num   4  4  1  1  2  1  4  2  2  4 ...
```

#verificando um resumo sobre o data frame

> summary(mtcars)

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

Primeiros passos com R

- Outros comandos úteis:

#limpar o conteúdo da tela

> CTRL + L

#acessar a documentação do R

> help()

#verificar histórico dos comandos

> history()

Tópicos

Introdução ao R

Primeiros passos com R

▶ **Análise de dados com R**

Análise de dados com R

Atividade

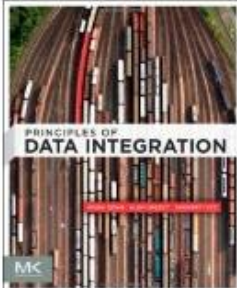
Análise de dados em sistemas de recomendação

Análise de dados com R

Quantas vezes você verificou/comprou itens recomendados por sites de e-commerce?

Related to Items You've Viewed

You viewed



Principles of Data Integration

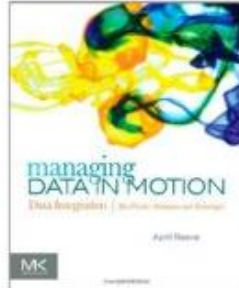
► AnHai Doan, Zachary Ives

★★★★☆ (2)

Hardcover: \$57.83

► [View or edit your browsing history](#)

Customers who viewed this also viewed



Managing Data in Motion: Data...

► April Reeve

★★★★★ (2)

Paperback: \$44.28

Kindle Edition: \$26.71



Connecting the Data: Data
Integration...

Angelo R. Bobak

★★★★★ (1)

Paperback: \$30.92



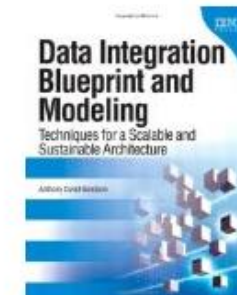
Data Matching: Concepts and...

► Peter Christen

★★★★☆ (6)

Hardcover: \$48.97

Kindle Edition: \$27.53



Data Integration Blueprint and...

► Anthony Giordano

★★★★☆ (7)

Hardcover

Análise de dados com R

Quantos vídeos você já assistiu que foram recomendados pelo YouTube?



Como parar de reclamar
por Arata Academy
46.826 visualizações
• 3 semanas atrás



Steven Tyler - Cryin' (Acoustic)
por Ismael Quispe
791.667 visualizações • 1 ano atrás



É possível ser feliz sendo pobre? - Flávio Gikovate
por Flávio Gikovate
14.361 visualizações • 1 ano atrás



O Teatro Mágico - Você me bagunça (Legendado)
por Monisa Segundo
1.608.098 visualizações • 4 anos atrás
[Mostrar mais](#)

Análise de dados com R

Você já encontrou indicações de vagas de emprego que fosse do seu interesse?



Vagas que podem ser de seu interesse



**Oportunidade essence -
Desenvolvedor Mobile**
essence. — São Paulo e
Região, Brasil

[Visualizar vaga](#)



Data Scientist
GetNinjas — Rebouças,
2472

[Visualizar vaga](#)



Java Developer
Experis — São Paulo Area,
Brazil

[Visualizar vaga](#)

Por que a recomendação é tão importante atualmente?

Análise de dados com R

- **Sistemas de recomendação:** área de analytics para gerar recomendações personalizadas a um usuário
- **Filtragem colaborativa:** tipo de sistema de recomendação para sugerir itens/produtos/serviços a partir de gostos similares de outros usuários

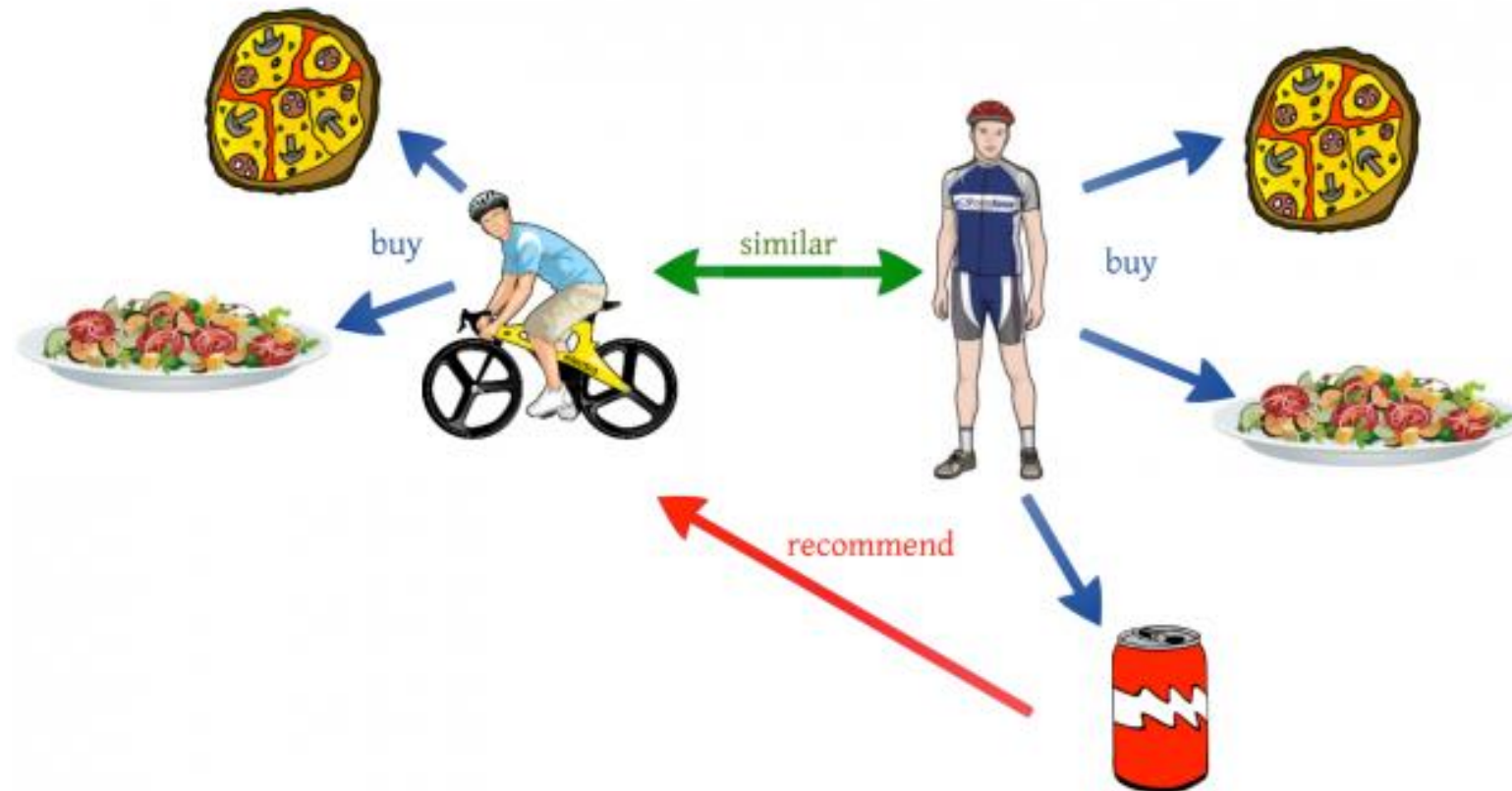


Análise de dados com R

Filtragem colaborativa item a item

- Similaridade entre itens i & j é computada isolando os usuários e aplicando uma técnica de cálculo de similaridade
- Recomenda os top-k-vizinhos mais próximos
- Recomendação é composta por itens que os usuários gostaram

Análise de dados com R



Análise de dados com R

Recomendação baseada no histórico de consumo de músicas dos usuários

The image is a screenshot of the Last.fm website. At the top is a red navigation bar with the 'last.fm' logo, a search bar labeled 'Pesquisar música', and links for 'Músicas', 'Ouvir', 'Eventos', and 'Tabelas'. On the right side of the bar are links for 'Associe-se' and 'Login'. Below the navigation bar, on the left, is a red cartoon character wearing headphones. To its right is the heading 'Descubra mais músicas' followed by a text box stating: 'A Last.fm é um serviço de descobertas de músicas que faz recomendações personalizadas com base nas músicas que você ouve.' Below this text is a search bar labeled 'Pesquisar artista, álbum ou faixa...'. To the right of the search bar is a red button that says 'Comece o seu perfil'. At the bottom of the page, there are four statistics, each with an icon and text: '75 bilhões de scrobbles' (headphones icon), '54 milhões de artistas' (star icon), '200 milhões de álbuns' (CD icon), and '640 milhões de faixas' (musical note icon).

last.fm Pesquisar música Músicas Ouvir Eventos Tabelas Associe-se Login

Descubra mais músicas

A Last.fm é um serviço de descobertas de músicas que faz recomendações personalizadas com base nas músicas que você ouve.

Pesquisar artista, álbum ou faixa...

Comece o seu perfil

75 bilhões de scrobbles 54 milhões de artistas 200 milhões de álbuns 640 milhões de faixas

Análise de dados com R

Exemplo



Legião Urbana

16.447.186 execuções (235.570 ouvintes)

Parecido com: Renato Russo, Cazuza, Capital Inicial, Os Paralamas Do Sucesso, Titãs

• rock

EM TOUR



Pitty

8.809.898 execuções (251.256 ouvintes)

Parecido com: Agridoce, Marjorie Estiano, Megh Stock, NX Zero, Luxúria

• rock

EM TOUR



Caetano Veloso

12.782.155 execuções (440.153 ouvintes)

Parecido com: Gal Costa, Gilberto Gil, Maria Bethânia, Chico Buarque, Tom Zé

• mpb

Análise de dados com R

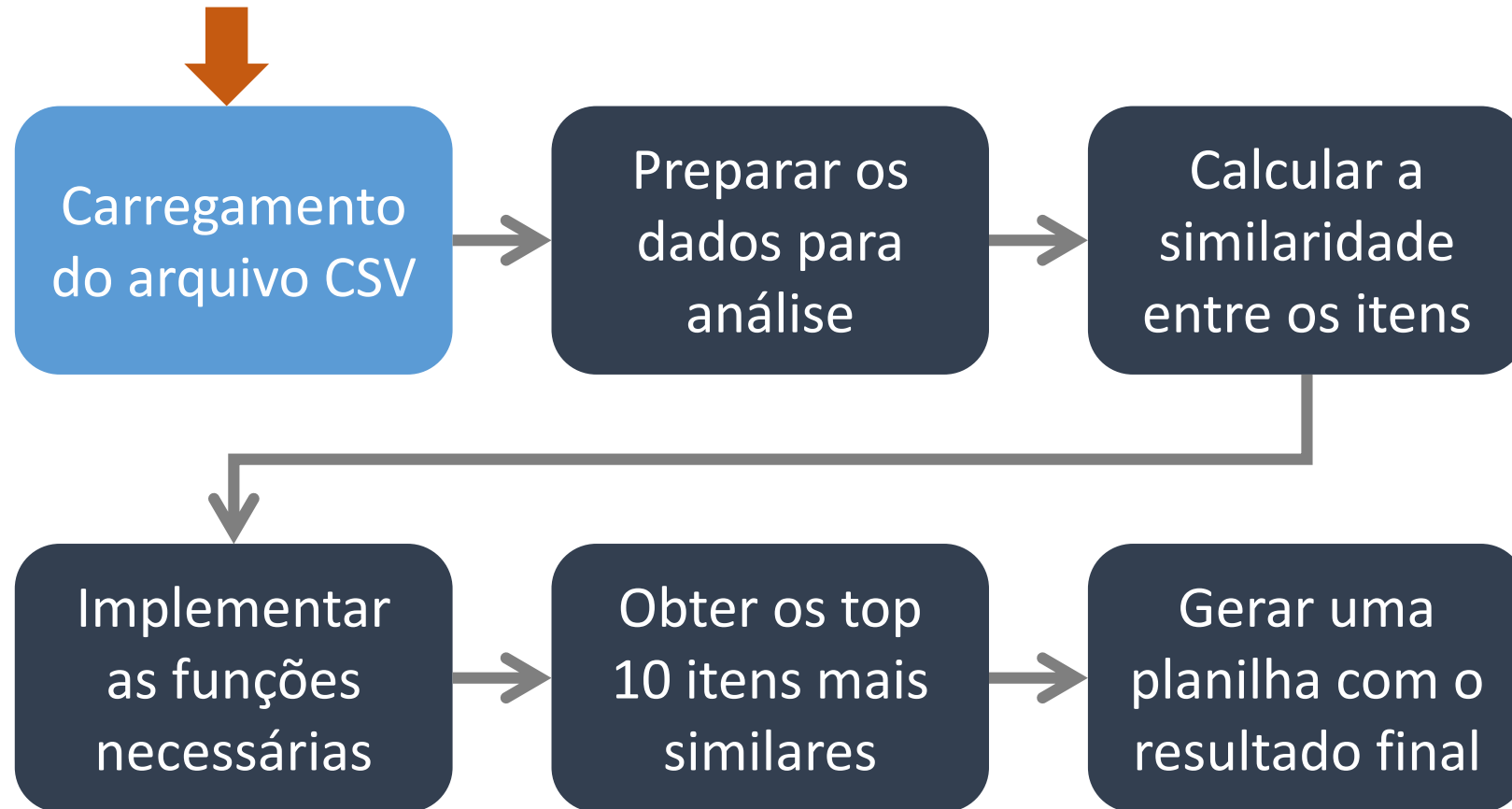
Base de dados: tabela usuário/artista com as preferências dos usuários da Last.fm

- Cada linha representa um usuário
- Cada coluna representa um artista
- Conteúdo da matriz indica as preferências de cada usuário

User, abba, ac.dc, adam green, aerosmith, afi, air

1,	0,	0,	0,	0,	0,	0
33,	0,	0,	1,	0,	0,	0
42,	0,	0,	0,	0,	0,	0
51,	0,	0,	0,	0,	0,	0
62,	0,	0,	0,	0,	0,	0
75,	0,	0,	0,	0,	0,	0

Análise de dados com R



Análise de dados com R

- Para padronizar o local de acesso, vamos criar um diretório chamado “Rnapratica”, dentro do diretório Documents, pelo Windows explorer.
- Salve o arquivo lastfm.csv dentro do diretório criado

Análise de dados com R

Configurando o diretório de trabalho

Console

```
> setwd("C:/Users/<nome_usuario>/Documents/Rnapratica")
```


Análise de dados com R

Verificar o diretório de trabalho atual

Console

```
> getwd()
```

```
[1] "C:/Users/Aluno/Documents/Rnapratica"
```

Análise de dados com R

Carregando o arquivo do tipo CSV no R

Console

```
> base <- read.csv(file="lastfm.csv")
```

Análise de dados com R

Visualizando a estrutura do data frame

Console

```
> str(base)
```

```
'data.frame': 1257 obs. of 286 variables:
```

```
$ user          : int  1 33 42 51 62 75 130 141 144 150 ...
```

```
$ a.perfect.circle : int  0 0 0 0 0 0 0 0 0 0 ...
```

```
$ abba          : int  0 0 0 0 0 0 0 0 0 0 ...
```

```
$ ac.dc         : int  0 0 0 0 0 0 0 0 0 0 ...
```

```
$ adam.green    : int  0 1 0 0 0 0 0 0 0 0 ...
```

```
...
```

Análise de dados com R

Visualizando os dados

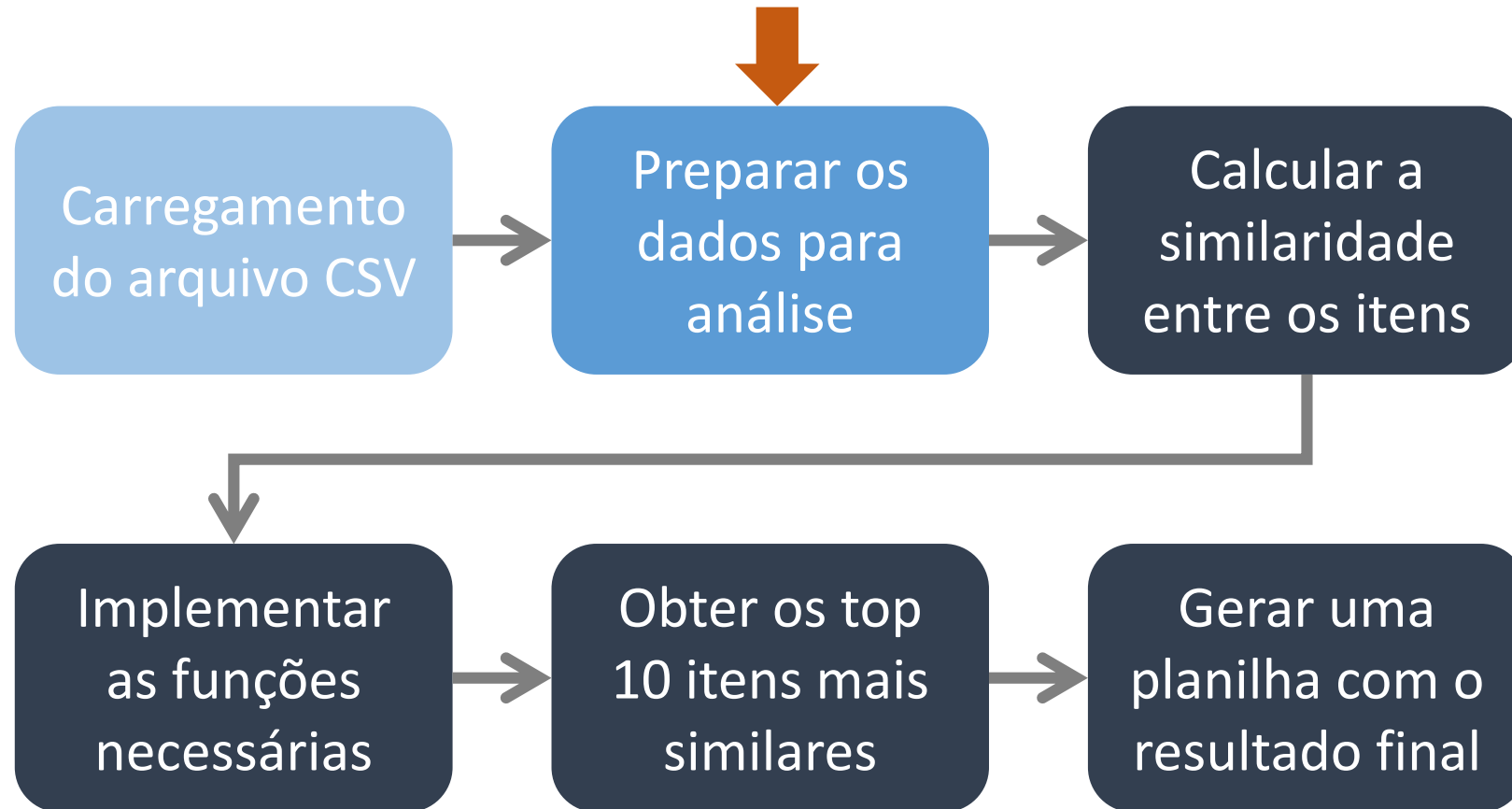
- Obtendo os resultados das primeiras 7 linhas e das colunas 1, 3, 4, 5, 6, 7 e 8

Console

```
> head(base[,c(1,3:8)])
```

```
  user abba ac.dc adam.green aerosmith afi air  
1    1  0    0    0    0    0    0  
2   33  0    0    1    0    0    0  
3   42  0    0    0    0    0    0  
4   51  0    0    0    0    0    0  
5   62  0    0    0    0    0    0  
6   75  0    0    0    0    0    0
```

Análise de dados com R



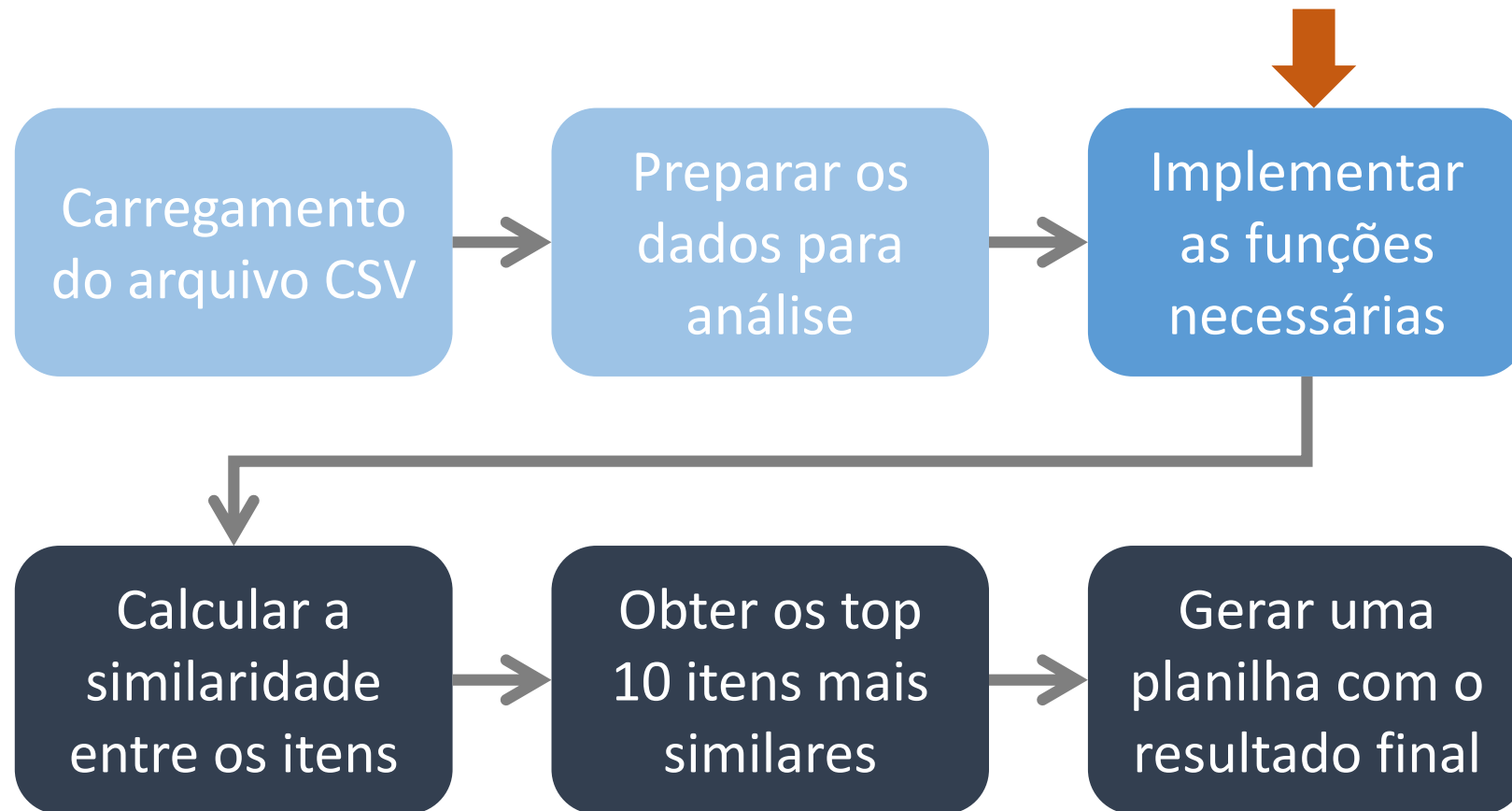
Análise de dados com R

Retirando a coluna de usuários do data frame

Console

```
> base.msc <- (base[,!(names(base) %in% c("user"))])
```

Análise de dados com R



Análise de dados com R

- Um dos principais problemas de mineração de dados está em encontrar similaridade entre objetos.
- Entre as medidas existentes, uma das adotadas é a **similaridade do cosseno**, que verifica o cosseno do ângulo entre dois vetores.

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} :$$

- O valor de similaridade varia entre 0 e 1, sendo que 0 indica que os itens não são similares, e 1 que são idênticos.

Análise de dados com R

Função para cálculo do cosseno do ângulo entre dois vetores

Exemplo:

$$\mathbf{d1} = (5,0,3,0,2,0,0,2,0,0)$$

$$\mathbf{d2} = (3,0,2,0,1,1,0,1,0,1)$$

Primeiro, calcule o produto vetorial dos pontos

$$\mathbf{d1} \times \mathbf{d2} = 5 \times 3 + 0 \times 0 + 3 \times 2 + 0 \times 0 + 2 \times 1 + 0 \times 1 + 0 \times 1 + 2 \times 1 + 0 \times 0 + 0 \times 1 = 25$$

Depois, calcule o módulo de d1 e de d2

$$|\mathbf{d1}| = \sqrt{5 \times 5 + 0 \times 0 + 3 \times 3 + 0 \times 0 + 2 \times 2 + 0 \times 0 + 0 \times 0 + 2 \times 2 + 0 \times 0 + 0 \times 0} = 6.481$$

$$|\mathbf{d2}| = \sqrt{3 \times 3 + 0 \times 0 + 2 \times 2 + 0 \times 0 + 1 \times 1 + 1 \times 1 + 0 \times 0 + 1 \times 1 + 0 \times 0 + 1 \times 1} = 4.12$$

$$\text{cosseno}(\mathbf{d1}, \mathbf{d2}) = 25 / (6.481 \times 4.12) = 0.94$$

Análise de dados com R

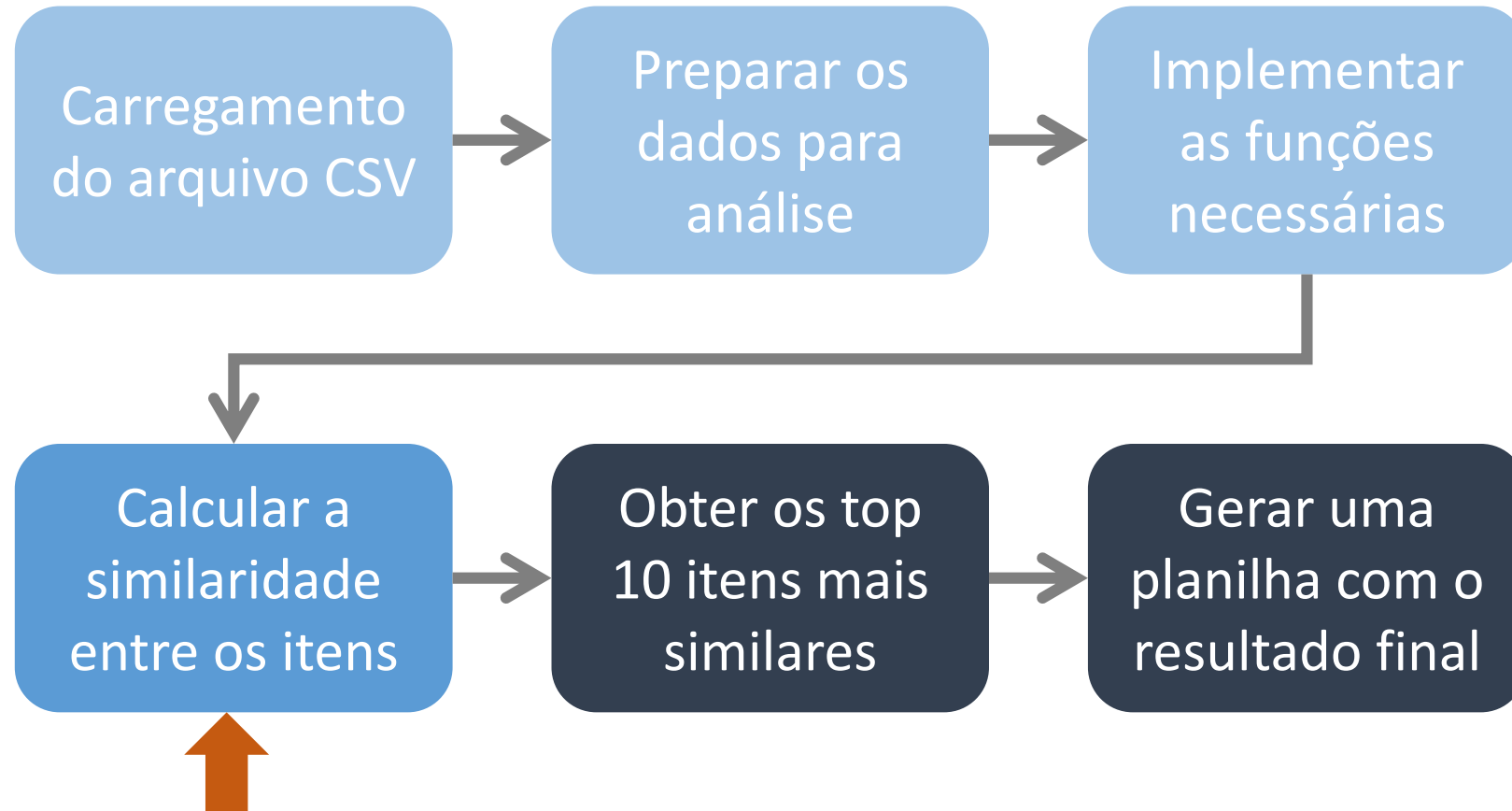
Função para cálculo do cosseno entre dois vetores

- Oferece uma medida de quão similares são dois itens

Console

```
> calculaCosseno<- function(x,y)
{
  cosseno <- sum(x*y) / (sqrt(sum(x*x)) * sqrt(sum(y*y)))
  return(cosseno)
}
```

Análise de dados com R



Análise de dados com R

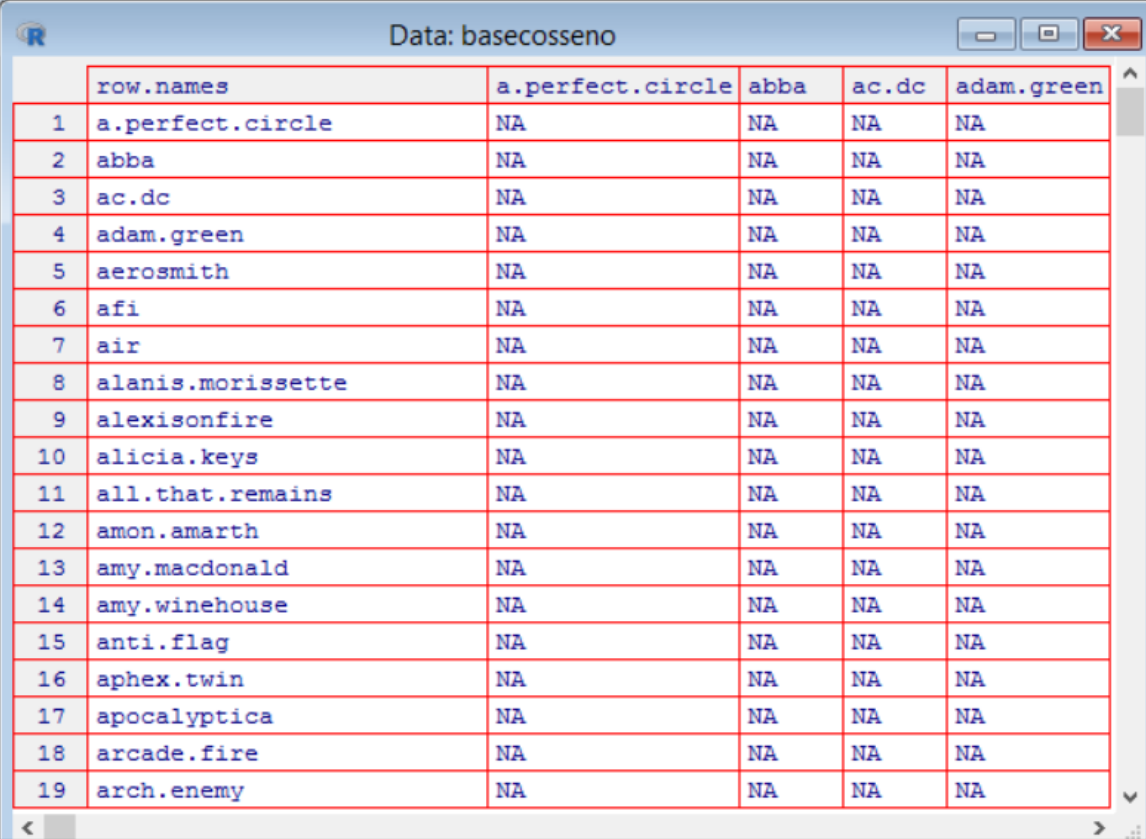
Criar uma matriz para armazenar as medidas de similaridade entre os itens

Console

```
> basecosseno <- matrix(NA, nrow=ncol(base), ncol=ncol(base),  
dimnames=list(colnames(base), colnames(base)))  
  
> View(basecosseno)
```

Análise de dados com R

Criar uma matriz para armazenar as medidas de similaridade entre os itens



The screenshot shows an R Data Viewer window titled "Data: basecosseno". It displays a matrix of similarity measures between 19 items. The items are listed in the first column, and the similarity measures are in the subsequent columns. The matrix is lower triangular, with the diagonal cells containing "NA".

	row.names	a.perfect.circle	abba	ac.dc	adam.green
1	a.perfect.circle	NA	NA	NA	NA
2	abba	NA	NA	NA	NA
3	ac.dc	NA	NA	NA	NA
4	adam.green	NA	NA	NA	NA
5	aerosmith	NA	NA	NA	NA
6	afi	NA	NA	NA	NA
7	air	NA	NA	NA	NA
8	alanis.morissette	NA	NA	NA	NA
9	alexisonfire	NA	NA	NA	NA
10	alicia.keys	NA	NA	NA	NA
11	all.that.remains	NA	NA	NA	NA
12	amon.amarth	NA	NA	NA	NA
13	amy.macdonald	NA	NA	NA	NA
14	amy.winehouse	NA	NA	NA	NA
15	anti.flag	NA	NA	NA	NA
16	aphex.twin	NA	NA	NA	NA
17	apocalyptica	NA	NA	NA	NA
18	arcade.fire	NA	NA	NA	NA
19	arch.enemy	NA	NA	NA	NA

Análise de dados com R

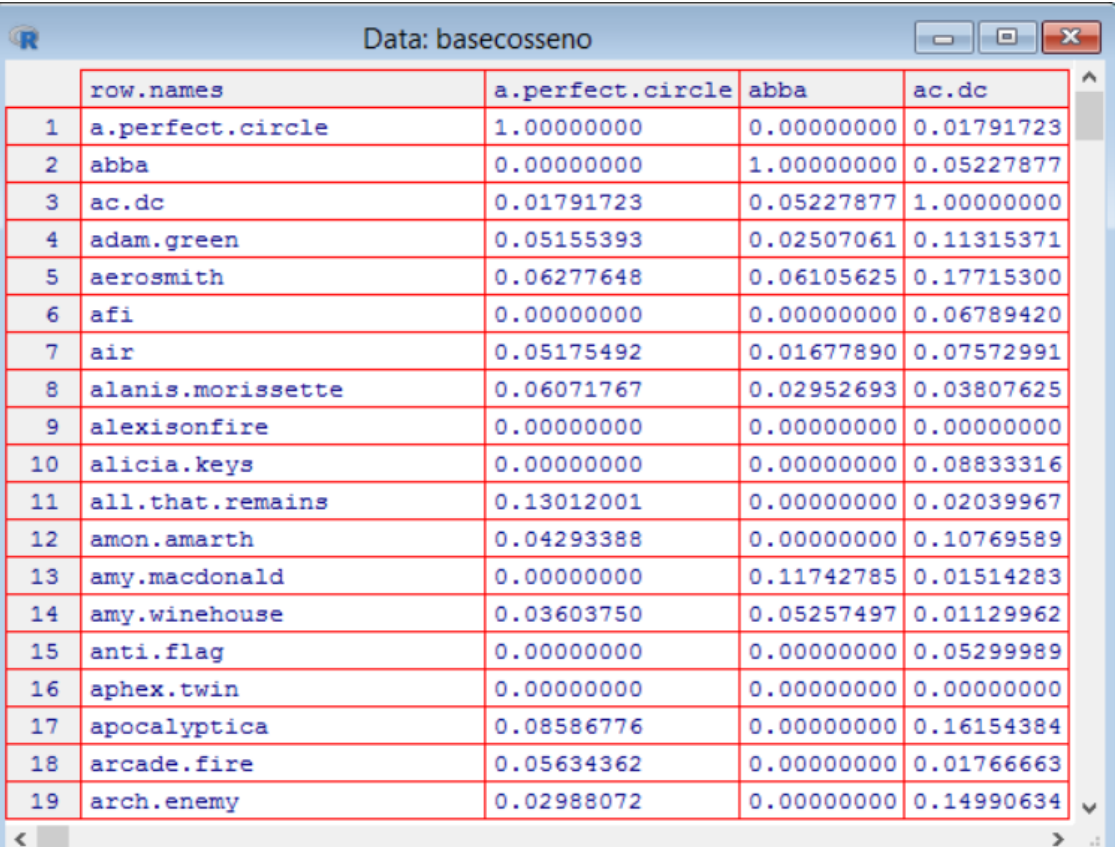
Calcular o cosseno de similaridade entre todas as colunas em uma matriz

Console

```
> for(i in 1:ncol(base)) {  
  for(j in 1:ncol(base)) {  
    basecosseno[i,j] <- calculaCosseno(as.matrix(base[i]), as.matrix(base[j]))  
  }  
}  
  
> View(basecosseno)
```

Análise de dados com R

Calcular o cosseno de similaridade entre todas as colunas em uma matriz



The screenshot shows an R window titled "Data: basecosseno". It displays a data frame with 19 rows and 5 columns. The first column contains row names, and the subsequent columns contain similarity scores for specific artists: 'a.perfect.circle', 'abba', and 'ac.dc'. The scores are calculated for all pairs of artists in the dataset.

	row.names	a.perfect.circle	abba	ac.dc
1	a.perfect.circle	1.00000000	0.00000000	0.01791723
2	abba	0.00000000	1.00000000	0.05227877
3	ac.dc	0.01791723	0.05227877	1.00000000
4	adam.green	0.05155393	0.02507061	0.11315371
5	aerosmith	0.06277648	0.06105625	0.17715300
6	afi	0.00000000	0.00000000	0.06789420
7	air	0.05175492	0.01677890	0.07572991
8	alanis.morissette	0.06071767	0.02952693	0.03807625
9	alexisonfire	0.00000000	0.00000000	0.00000000
10	alicia.keys	0.00000000	0.00000000	0.08833316
11	all.that.remains	0.13012001	0.00000000	0.02039967
12	amon.amarth	0.04293388	0.00000000	0.10769589
13	amy.macdonald	0.00000000	0.11742785	0.01514283
14	amy.winehouse	0.03603750	0.05257497	0.01129962
15	anti.flag	0.00000000	0.00000000	0.05299989
16	aphex.twin	0.00000000	0.00000000	0.00000000
17	apocalyptica	0.08586776	0.00000000	0.16154384
18	arcade.fire	0.05634362	0.00000000	0.01766663
19	arch.enemy	0.02988072	0.00000000	0.14990634

Análise de dados com R

Converter a matriz de similaridade em um data frame

Console

```
> basecos seno <- as.data.frame(basecos seno)
```


Análise de dados com R

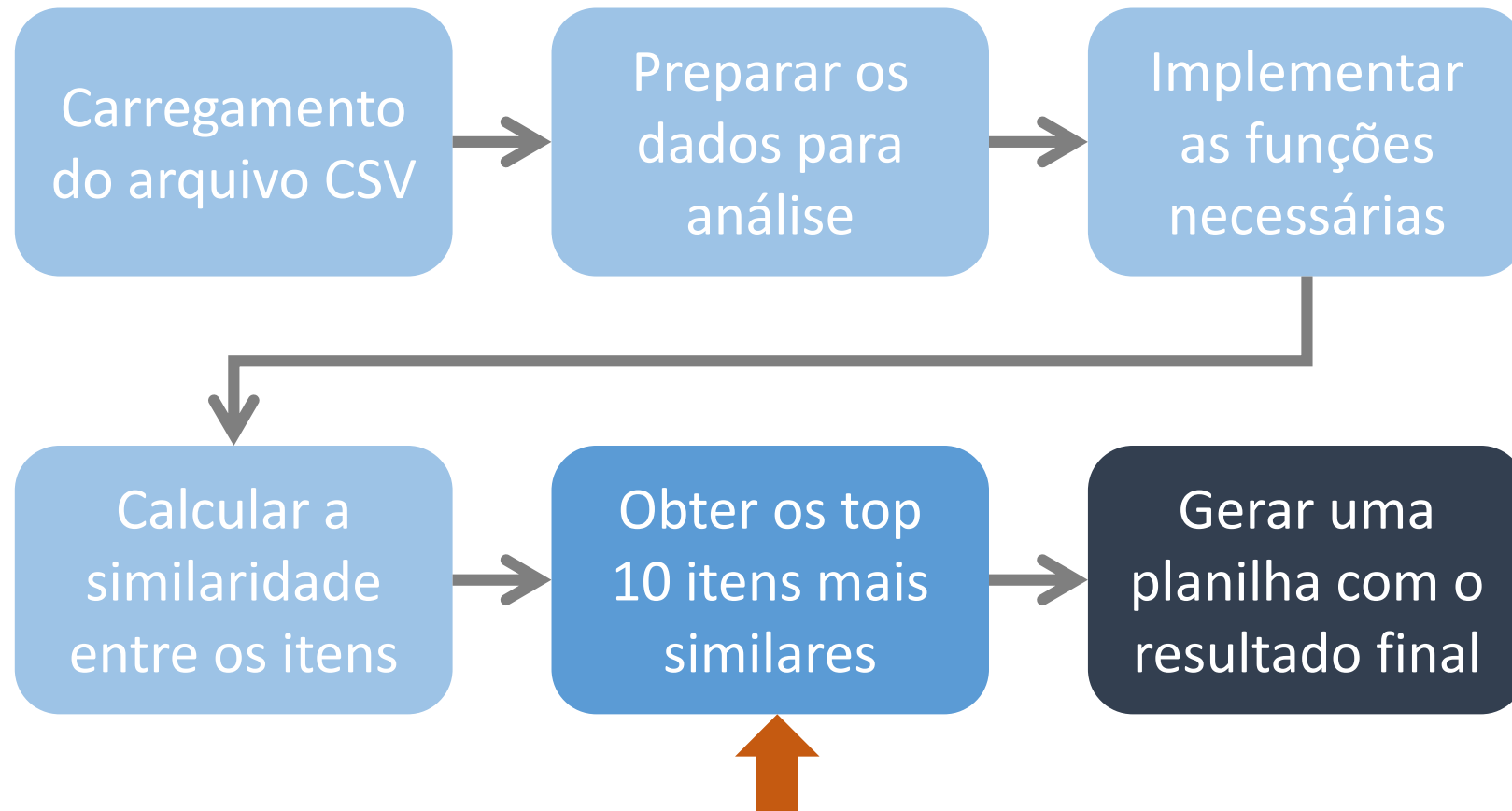
Visualizar o conteúdo do data frame

Console

```
> head(basecosseno[,c(1,2:5)])
```

	a.perfect.circle	abba	ac.dc	adam.green	aerosmith
a.perfect.circle	1.00000000	0.00000000	0.01791723	0.05155393	0.06277648
abba	0.00000000	1.00000000	0.05227877	0.02507061	0.06105625
ac.dc	0.01791723	0.05227877	1.00000000	0.11315371	0.17715300
adam.green	0.05155393	0.02507061	0.11315371	1.00000000	0.05663655
aerosmith	0.06277648	0.06105625	0.17715300	0.05663655	1.00000000
afi	0.00000000	0.00000000	0.06789420	0.00000000	0.00000000

Análise de dados com R



Análise de dados com R

Criar uma nova matriz para armazenar os top-10 vizinhos mais próximos de cada item

Console

```
> basevizinhos <- matrix(NA, nrow=ncol(basecosseno), ncol=11,  
dimnames=list(colnames(basecosseno)))
```

Análise de dados com R

Ordenar a matriz em ordem decrescente de similaridade

Console

```
> for(i in 1:ncol(base)) {  
  basevizinhos[i,] <- (t(head(n=11, rownames(basecosseno[order(  
basecosseno[,i], decreasing=TRUE),][i]))));  
}
```

Análise de dados com R

Visualizar a matriz

Console

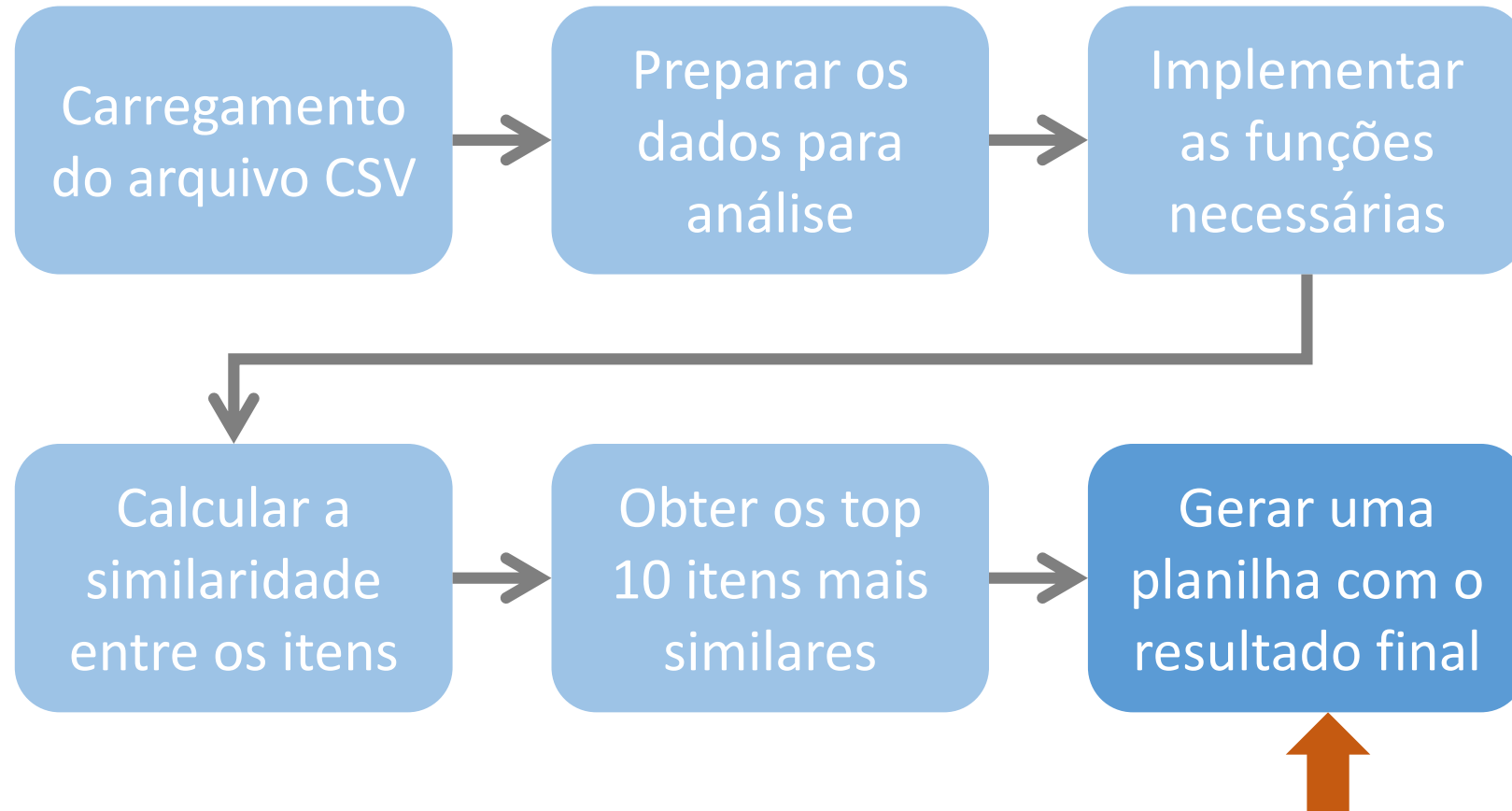
```
> View(basevizinhos)
```

Análise de dados com R

Visualizar a matriz

Data: basevizinhos					
	row.names	V1	V2	V3	V4
1	a.perfect.circle	a.perfect.circle	tool	dredg	deftones
2	abba	abba	madonna	robbie.williams	elvis.presley
3	ac.dc	ac.dc	red.hot.chili.peppers	metallica	iron.maiden
4	adam.green	adam.green	the.libertines	the.strokes	babyshambles
5	aerosmith	aerosmith	u2	led.zepplin	metallica
6	afi	afi	funeral.for.a.friend	rise.against	fall.out.boy
7	air	air	massive.attack	goldfrapp	morceeba
8	alanis.morissette	alanis.morissette	tori.amos	alicia.keys	red.hot.chili.peppers
9	alexisonfire	alexisonfire	atreyu	underoath	funeral.for.a.friend
10	alicia.keys	alicia.keys	beyonce	norah.jones	maria.mena
11	all.that.remains	all.that.remains	heaven.shall.burn	as.i.lay.dying	parkway.drive
12	amon.amarth	amon.amarth	dark.tranquillity	equilibrium	eluveitie
13	amy.macdonald	amy.macdonald	maria.mena	lily.allen	leona.lewis
14	amy.winehouse	amy.winehouse	norah.jones	duffy	jack.johnson
15	anti.flag	anti.flag	rise.against	nofx	millencolin
16	aphex.twin	aphex.twin	bjork	portishead	boards.of.canada
17	apocalyptica	apocalyptica	subway.to.sally	in.extremo	metallica
18	arcade.fire	arcade.fire	the.shins	the.national	belle.and.sebastian
19	arch.enemy	arch.enemy	in.flames	children.of.bodom	slayer

Análise de dados com R



Análise de dados com R

Gravar o resultado em um arquivo csv

Console

```
> write.csv(file="top10lastfm.csv",x=basevizinhos[,-1])
```


Análise de dados com R

Abra o arquivo e visualizar o resultado

1	Column1	Column2	Column3	Column4	Column5	Column6	Column7	Column8	Column9	Column10
2		V1	V2	V3	V4	V5	V6	V7	V8	V9
3	a.perfect.circle	tool	dredg	deftones	porcupine.tree	nine.inch.nails	incubus	system.of.a.down	opeth	the.smash
4	abba	madonna	robbie.williams	elvis.presley	michael.jackson	queen	the.beatles	kelly.clarkson	groove.coverage	duffy
5	ac.dc	red.hot.chili.peppers	metallica	iron.maiden	the.offspring	black.sabbath	die.toten.hosen	rammstein	judas.priest	the.beatles
6	adam.green	the.libertines	the.strokes	babyshambles	radiohead	franz.ferdinand	the.kooks	foo.fighters	the.white.stripes	the.beatles
7	aerosmith	u2	led.zepelin	metallica	ac.dc	lenny.kravitz	the.rolling.stones	jack.johnson	red.hot.chili.peppers	robbie.williams
8	afi	funeral.for.a.friend	rise.against	fall.out.boy	anti.flag	sum.41	billy.talent	lostprophets	silverstein	millencolin
9	air	massive.attack	goldfrapp	morceeba	thievery.corporation	jamiroquai	nouvelle.vague	coldplay	portishead	daft.punk
10	alanis.morissette	tori.amos	alicia.keys	red.hot.chili.peppers	kelly.clarkson	dido	coldplay	pearl.jam	jack.johnson	norah.jones
11	alexisonfire	atreyu	underoath	funeral.for.a.friend	silverstein	killswitch.engine	rise.against	caliban	enter.shikari	three.days
12	alicia.keys	beyonce	norah.jones	maria.mena	black.eyed.peas	lenny.kravitz	amy.winehouse	christina.aguilera	rihanna	duffy
13	all.that.remains	heaven.shall.burn	as.i.lay.dying	parkway.drive	trivium	caliban	killswitch.engine	chimaira	atreyu	bullet.for.r
14	amon.amarth	dark.tranquillity	equilibrium	eluveitie	ensiferum	in.flames	finntroll	die.apokalyptischen.reiter	dimmu.borgir	children.of
15	amy.macdonald	maria.mena	lily.allen	leona.lewis	amy.winehouse	jason.mraz	coldplay	james.morrison	avril.lavigne	keane
16	amy.winehouse	norah.jones	duffy	jack.johnson	kate.nash	lily.allen	the.kooks	leona.lewis	the.killers	alicia.keys
17	anti.flag	rise.against	nofx	millencolin	bad.religion	dropkick.murphys	misfits	the.hives	flogging.molly	die.toten.h
18	aphex.twin	bjork	portishead	boards.of.canada	elliott.smith	cocorosie	beirut	radiohead	the.clash	the.doors
19	apocalyptic	subway.to.sally	in.extremo	metallica	rammstein	nightwish	manowar	disturbed	iron.maiden	arch.enem
20	arcade.fire	the.shins	the.national	belle.and.sebastian	the.decemberists	sufjan.stevens	cat.power	stars	interpol	bloc.party
21	arch.enemy	in.flames	children.of.bodom	slayer	amon.amarth	iron.maiden	kreator	ensiferum	dimmu.borgir	hammerfa
22	arctic.monkeys	the.kooks	bloc.party	the.killers	mando.diao	muse	franz.ferdinand	the.white.stripes	the.hives	beatsteaks
23	as.i.lay.dying	all.that.remains	killswitch.engine	heaven.shall.burn	caliban	parkway.drive	atreyu	bring.me.the.horizon	trivium	chimaira
24	atb	scooter	groove.coverage	faithless	nelly.furtado	rihanna	daft.punk	bloodhound.gang	the.prodigy	cascada
25	atreyu	underoath	silverstein	as.i.lay.dying	killswitch.engine	alexisonfire	caliban	the.used	trivium	all.that.rer
26	audioslave	pearl.jam	rage.against.the.machine	foo.fighters	tenacious.d	seether	linkin.park	stone.sour	red.hot.chili.peppers	the.white.s
27	avril.lavigne	simple.plan	evanescence	kelly.clarkson	linkin.park	pink	christina.aguilera	leona.lewis	lady.gaga	sum.41
28	babyshambles	the.libertines	arctic.monkeys	the.kooks	blur	death.cab.for.cutie	the.strokes	franz.ferdinand	adam.green	bright.eyes
29	bad.religion	nofx	anti.flag	millencolin	rise.against	die.toten.hosen	dropkick.murphys	flogging.molly	red.hot.chili.peppers	the.offspri

Considerações finais

Análise de dados com R

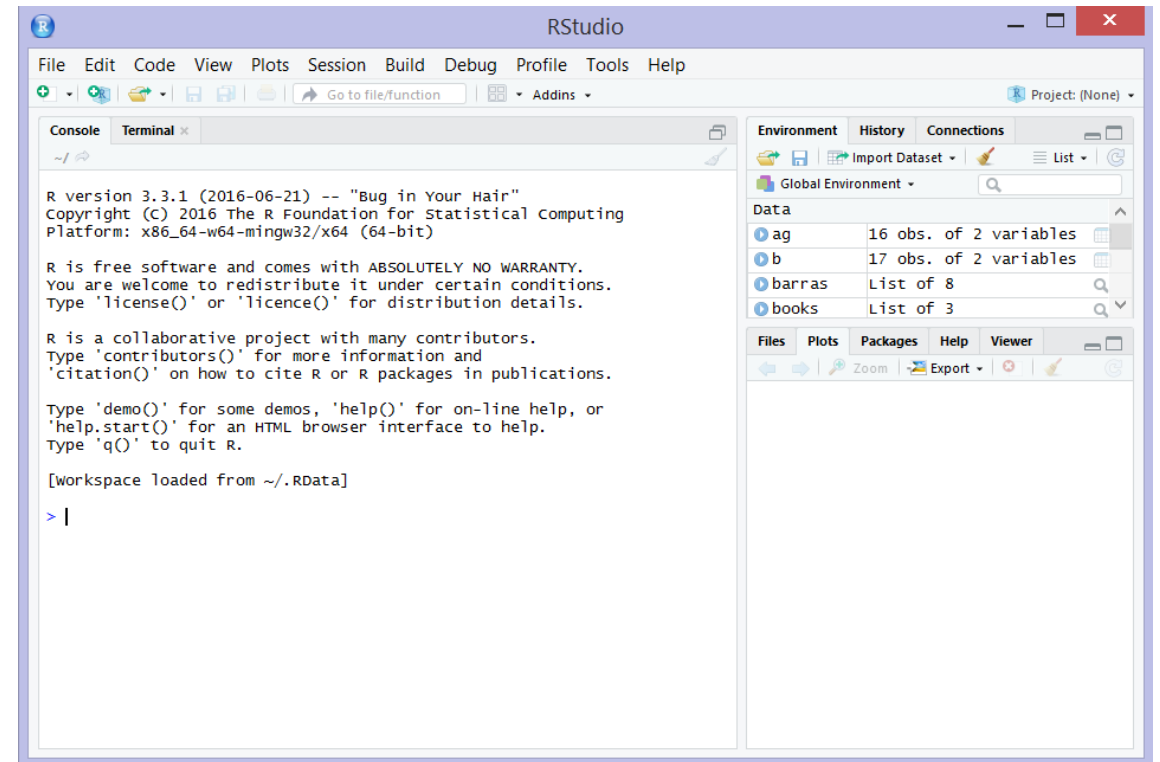
Considerações finais

- A atividade apresentada é um pequeno exemplo de uma visão inicial dos comandos utilizados por cientistas de dados para analisar dados, com o objetivo de gerar recomendações.
- Em projetos de análises de dados, no entanto, conforme identificado na literatura, é comum a execução de um conjunto mais extenso de comandos e manipulações de dados, de acordo com o objetivo pretendido.

Análise de dados com R

Sugestão de ferramenta

- RStudio - <https://www.rstudio.com/products/rstudio/download/>
- IDE com recursos adicionais para análise de dados utilizando a linguagem R



Um **BIG** Obrigada!

Referências

- Amatriain, X., Jaimes, A., Oliver, N., and Pujol, J. M. (2011). **Data mining methods for recommender systems**. In Recommender systems handbook, pages 39–71. Springer.
- PROVOST, Foster; FAWCETT, Tom. **Data Science for Business: What you need to know about data mining and data-analytic thinking**. " O'Reilly Media, Inc.", 2013.
- ABEDIN, Jaynal. **Data Manipulation with R**. Apress, 2018.
- FISCHETTI, Tony. **Data Analysis with R**. Packt Publishing Ltd, 2015.
- Ekstrand, M. D., Riedl, J. T., Konstan, J. A., et al. (2011). **Collaborative filtering recommender systems**. Foundations and Trends R in Human–Computer Interaction, 4(2):81–173.
- Forte, R. M. (2015). **Mastering predictive analytics with R**. Packt Publishing Ltd.
- Velleman, P. F. and Hoaglin, D. C. (1981). **Applications, basics, and computing of exploratory data analysis**. Duxbury Press.

Sobre os autores



Rosângela de Fátima Pereira Marquesone Pesquisadora no Laboratório de Arquitetura e Redes de Computadores (LARC-USP), atuando nas áreas de computação em nuvem e Big Data. Atua como professora e palestrante de cursos de Big Data para empresas e programas de MBA, tendo ministrado mais de 500 horas de aula sobre o tema. Também atua como revisora de código no Nanodegree Analista de Dados da rede de cursos on-line Udacity. É Mestre e doutoranda em Engenharia de Computação pela Escola Politécnica da Universidade de São Paulo (EPUSP). É Bacharel em Administração de Empresas pela Universidade Estadual do Norte do Paraná (UENP) (2007), Tecnóloga em Análise e Desenvolvimento de Sistemas pela Universidade Tecnológica Federal do Paraná (UTFPR) (2011) e Especialista em Tecnologia Java pela UTFPR (2010). É autora do livro “Big Data – Técnicas e tecnologias para extração de valor dos dados”, publicado pela editora Casa do Código. Seus principais interesses de pesquisa são: Big Data, computação em nuvem e people analytics. Também se interessa por temas como design thinking, mulheres na tecnologia e empreendedorismo social.

Sobre os autores



Francisco Pereira Junior possui graduação em Tecnologia em Processamento de Dados pelo Centro de Estudos Superiores de Londrina (1998) e mestrado em Ciência da Computação pela Universidade Estadual de Maringá (2006). É professor efetivo do Departamento Acadêmico de Computação da Universidade Tecnológica Federal do Paraná (UTFPR) atuando em cursos de graduação e pós-graduação (Engenharia de Computação, Engenharia de Software, Análise e Desenvolvimento de Sistemas, Tecnologia Java e Informática Aplicada à Educação). Também é representante Institucional da Universidade Tecnológica Federal do Paraná Câmpus Cornélio Procópio (UTFPR/CP) junto a Sociedade Brasileira de Computação (SBC). Participa de projetos e tem interesse em pesquisas com ênfase em Processamento de Alto Desempenho (Cluster / Grid / Nuvem / Programação Paralela - MPI), Big Data, Hadoop e todo seu ecossistema.

Sobre os autores



Tereza Cristina Melo de Brito Carvalho Graduada em 1980 em Engenharia Eletrônica, em 1988 como mestre e em 1996 como doutora na área de redes de computadores pela Escola Politécnica da USP (Poli). Concluiu o Sloan Fellows Program em 2002 pelo MIT – Massachusetts Institute of Technology, Boston – EUA. Já trabalhou na Siemens, Nuremberg – Alemanha, e na France Telecom, Every – França. Recebeu diversos prêmios, como: Prêmio InfoExame Inovação em Iniciativa Verde (2010), Prêmio e Menção Honrosa do Prêmio Governador Mário Covas em Inovação (2009 e 2008) do Governo de Estado de São Paulo pelo projeto do CEDIR (Centro de Descarte e Reuso de Resíduos de Informática) e de Criação do Selo Verde da USP, Personalidade em Tecnologia pela InfoExame (2005 e 2007) e Executiva em TI pela ABACO (2006). Foi diretora do CCE (Centro de Computação Eletrônica) da USP de 2006-2010. Atualmente, é Assessora para Projetos Especiais da CTI (Coordenadoria de Tecnologia de Informação) da USP, coordenadora do CEDIR, coordenadora geral do LASSU (Laboratório de Sustentabilidade em TIC) e membro do conselho diretor do LARC (Laboratório de Arquitetura e Redes de Computadores), ambos laboratórios de pesquisa do PCS (Departamento de Engenharia de Computação e Sistemas Digitais) da Escola Politécnica da USP. É professora assistente do PCS/Poli. Atua em projetos de pesquisa e desenvolvimento nas áreas de Sistemas de Informação, redes de comunicação, gerenciamento, segurança, Governança e Sustentabilidade em TIC. Possui mais de 100 artigos científicos e tecnológicos publicados em revistas indexadas, conferências internacionais e nacionais.