



Escola Politécnica de Pernambuco
Especialização em Ciência de Dados e Analytics

Estatística Computacional

Aula 1.2 – Aplicações Computacionais da Estatística – PARTE I

Prof. Dr. Rodrigo Lins Rodrigues

rodrigolins.rodrigues@ufrpe.br

Conteúdo programático

- ✓ Conhecendo softwares estatísticos;
- ✓ Porque cientistas de dados utilizam R;
- ✓ Conhecendo a linguagem;
- ✓ Tipos de dados;
- ✓ Trabalhando com funções;
- ✓ Importação de bases de dados;
- ✓ Plotagem de gráficos;
- ✓ Sumarização descritiva de dados;
- ✓ Correlação
- ✓ Projeto prático.



Recursos necessários

- ✓ Laboratório de Informática;
- ✓ Computadores com R basic instalado
- ✓ Computadores com RStudio instalado
- ✓ Acesso a internet para instalação de pacotes;





Conhecendo Softwares Estatísticos

Softwares Estatísticos

- Conhecendo softwares estatísticos;
 - ✓ Atualmente existem dezenas de softwares estatísticos;
 - ✓ É praticamente impossível, imaginar “a vida” de um **analista de dados** sem os recursos computacionais atuais;
 - ✓ Um cientista de dados deve conhecer o máximo de softwares de análises de dados;



Software SPSS

- Conhecendo softwares estatísticos;
 - ✓ Acrônimo de *Statistical Package for the Social Science*;
 - ✓ **Software estatístico** para as ciências sociais;
 - ✓ Teve a sua primeira **versão em 1968** e é um dos programas de análise estatística mais usados nas ciências sociais;
 - ✓ É um software **proprietário**, atualmente pela IBM.



Software SPSS

*Employee data.sav [DataSet1] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Graphs Custom Utilities Add-ons Window Help

Visible: 10 of 10 Variables

	id	gender	bdate	educ	jobcat	salary	salbegin	jobtime	p
1	1	Male	02/03/1952	15	Manager	\$57,000	\$27,000	98	
2	2	Male	05/23/1958	16	Clerical	\$40,200	\$18,750	98	
3	3	Female	07/26/1929	12	Clerical	\$21,450	\$12,000	98	
4	4	Female	04/15/1947	8	Clerical	\$21,900	\$13,200	98	
5	5	Male	02/09/1955	15	Clerical	\$45,000	\$21,000	98	
6	6	Male	08/22/1958	15	Clerical	\$32,100	\$13,500	98	
7	7	Male	04/26/1956	15	Clerical	\$36,000	\$18,750	98	
8	8	Female	05/06/1966	12	Clerical	\$21,900	\$9,750	98	
9	9	Female	01/23/1946	15	Clerical	\$27,900	\$12,750	98	
10	10	Female	02/13/1946	12	Clerical	\$24,000	\$13,500	98	
11	11	Female	02/07/1950	16	Clerical	\$30,300	\$16,500	98	
12	12	Male	01/11/1966	8	Clerical	\$28,350	\$12,000	98	
13	13	Male	07/17/1960	15	Clerical	\$27,750	\$14,250	98	
14	14	Female	02/26/1949	15	Clerical	\$35,100	\$16,800	98	

Data View Variable View

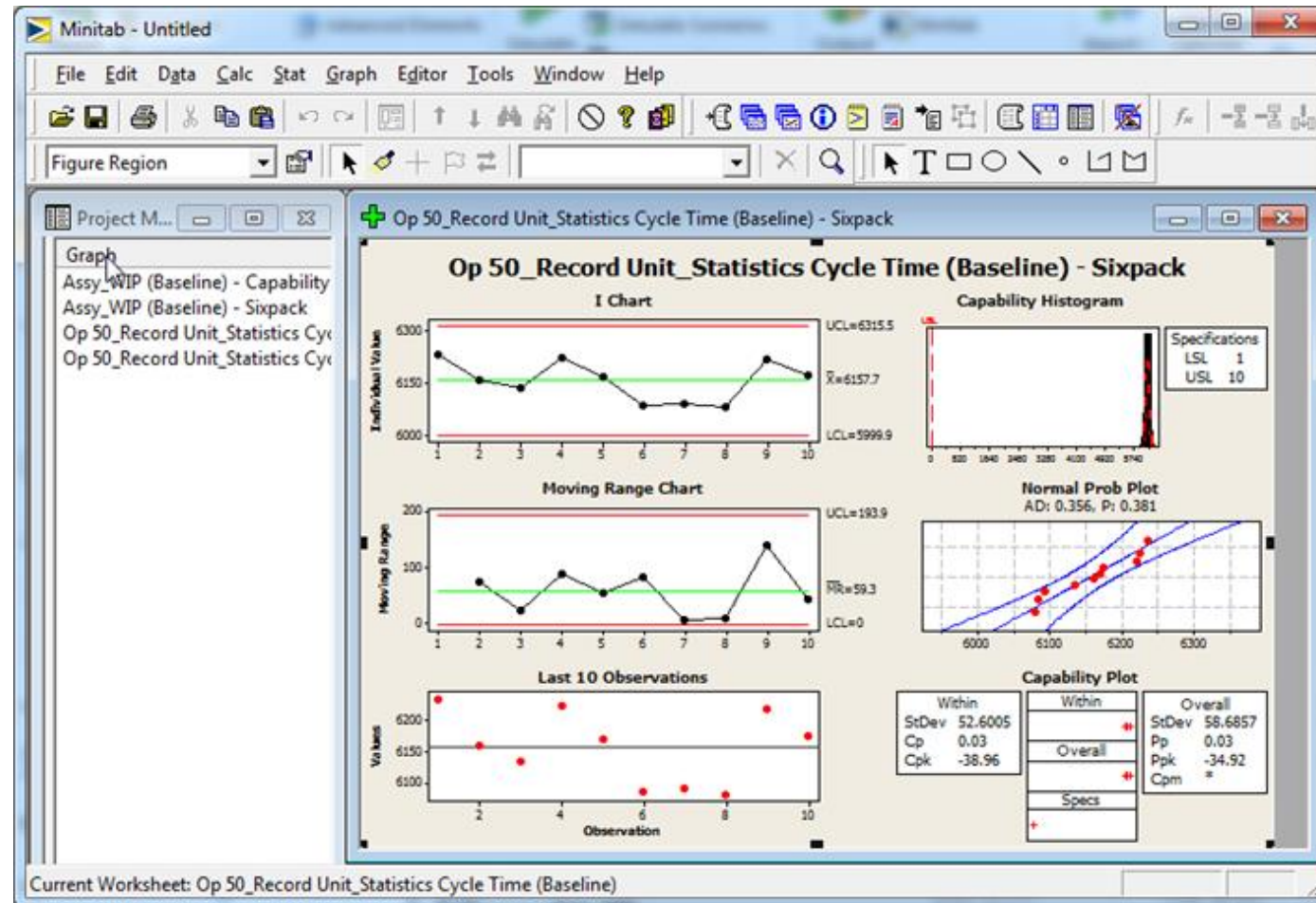
IBM SPSS Statistics Processor is ready Unicode:ON

Software Minitab

- Conhecendo softwares estatísticos;
 - ✓ É muito utilizado nas universidades nos cursos introdutórios de estatística;
 - ✓ Foi desenvolvido em 1972;
 - ✓ Fácil de usar e de aprender;
 - ✓ Possui versão grátis para estudantes.



Software Minitab

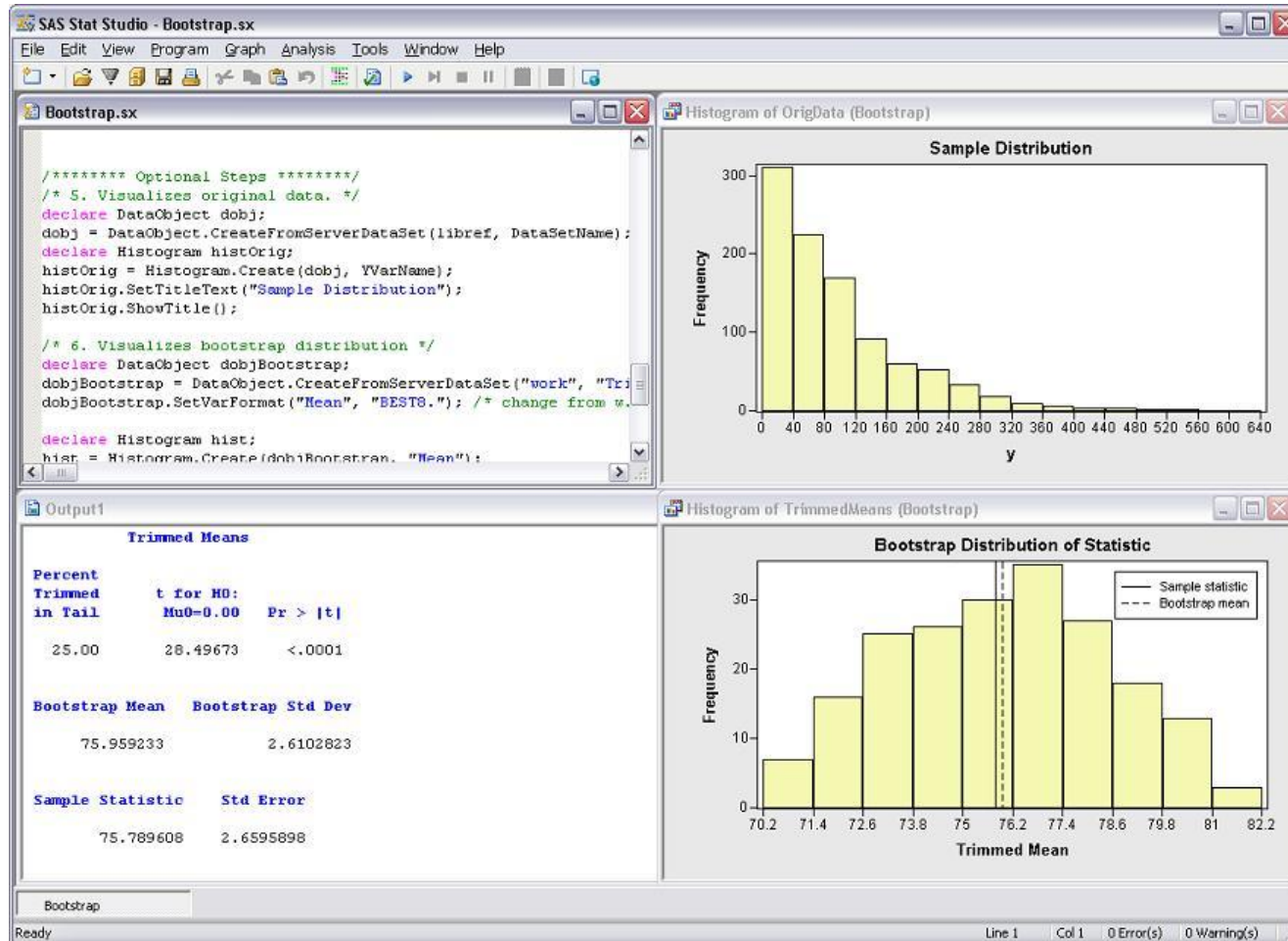


Software SAS

- Conhecendo softwares estatísticos;
 - ✓ Derivado do nome "Statistical Analysis System;
 - ✓ Muito utilizado nas ciências agronômicas;
 - ✓ Tem uma linguagem própria de script;
 - ✓ Possui versão grátis para estudantes.



Software SAS

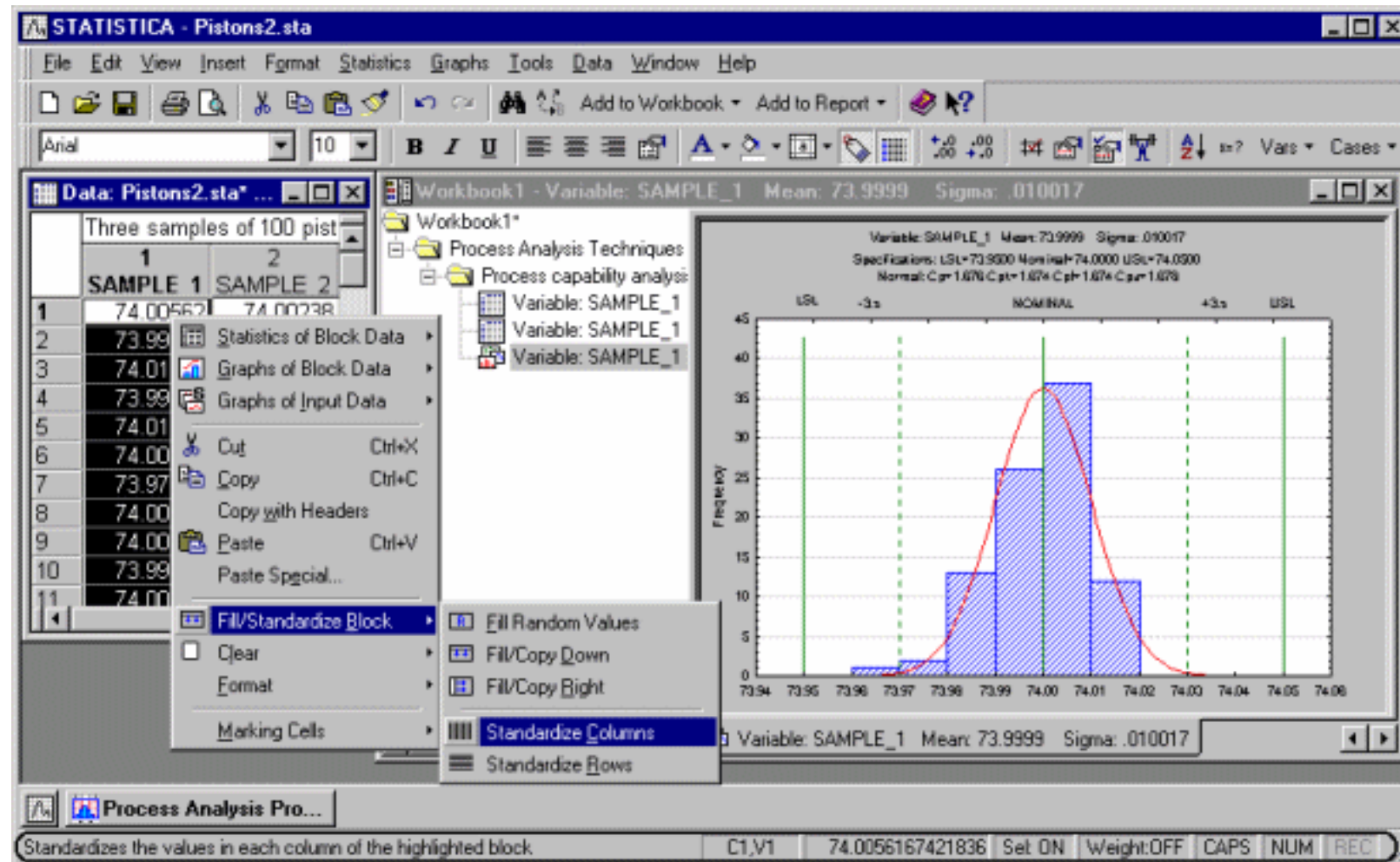


Software Statistica

- Conhecendo softwares estatísticos;
 - ✓ Um dos softwares estatísticos mais fáceis de ser manipulado;
 - ✓ Muito utilizado por engenheiros químicos;
 - ✓ Possui versão grátis para estudantes.



Software Statistica



Software KNIME

- Conhecendo softwares estatísticos;
 - ✓ Apesar de realizar análises estatísticas, é mais voltado pra mineração de dados;
 - ✓ Possibilita *workflow* de análise;
 - ✓ É possível integrar o processo de ETL com a análise em um único fluxo.



Software KNIME

The screenshot displays the KNIME software interface with a workflow diagram in the center. The workflow consists of the following nodes and connections:

- Time Generator** (Node 41) connects to **Table Creator** (Node 38).
- Table Creator** (Node 38) connects to **Generic Loop Start** (Node 35).
- Generic Loop Start** (Node 35) connects to **Table Row To Variable Loop Start** (Node 30).
- Table Row To Variable Loop Start** (Node 30) connects to **Time Difference** (Node 40).
- Time Difference** (Node 40) connects to **Twitter Search** (Node 2).
- Twitter Search** (Node 2) connects to **Twitter API Connector** (Node 9).
- Twitter API Connector** (Node 9) connects to **Twitter Users** (Node 3).
- Twitter Users** (Node 3) connects to **Loop End** (Node 25).
- Loop End** (Node 25) connects to **CSV Writer** (Node 32).

The left sidebar contains the **KNIME Explorer** and **Node Repository**. The **Node Repository** shows a tree structure under **Time Series** with various nodes like **Seasonality Correction**, **Date Field Extractor**, and **Time Generator**.

The bottom right pane shows the **KNIME Console** with the following output:

```
Execute failed: 429:Returned in API v1.1 when a request cannot be served due to the application's rate l
Node created an empty data table.
p End
No variable selected
p End
No variable selected
```

Software Rapidminer

- Conhecendo softwares estatísticos:
 - ✓ Muito utilizado na área de mineração de textos;
 - ✓ Tem uma interface muito amigável;
 - ✓ Tem uma versão grátis mas a versão completa é paga.



Software Rapidminer

The screenshot displays the RapidMiner software interface with a workflow titled "Main Process". The workflow consists of the following steps:

- Retrieve**: Connects to a data source.
- Normalization**: Normalizes the data.
- MissingValue...**: Handles missing values.
- LibSVM**: The core classification model.
- Nominal2Bino...**: Converts nominal attributes to binary.
- Nominal2Num...**: Converts nominal attributes to numeric.

The workflow is connected to a data source named "data" (none) in the "Samples" repository. The "LibSVM" process is configured with the following parameters:

- svm type: C-SVC
- kernel type: rbf
- gamma: 22644346174132
- C: 85795883818439
- epsilon: 0.0010
- ☐ calculate confidences

The "Problems" panel at the bottom shows two potential problems:

Message	Fixes	Location
Attribute filter does not match any attributes.	Select all attributes.	Nominal2Binomin...
Attribute filter does not match any attributes.	Select all attributes.	Nominal2Numeric...

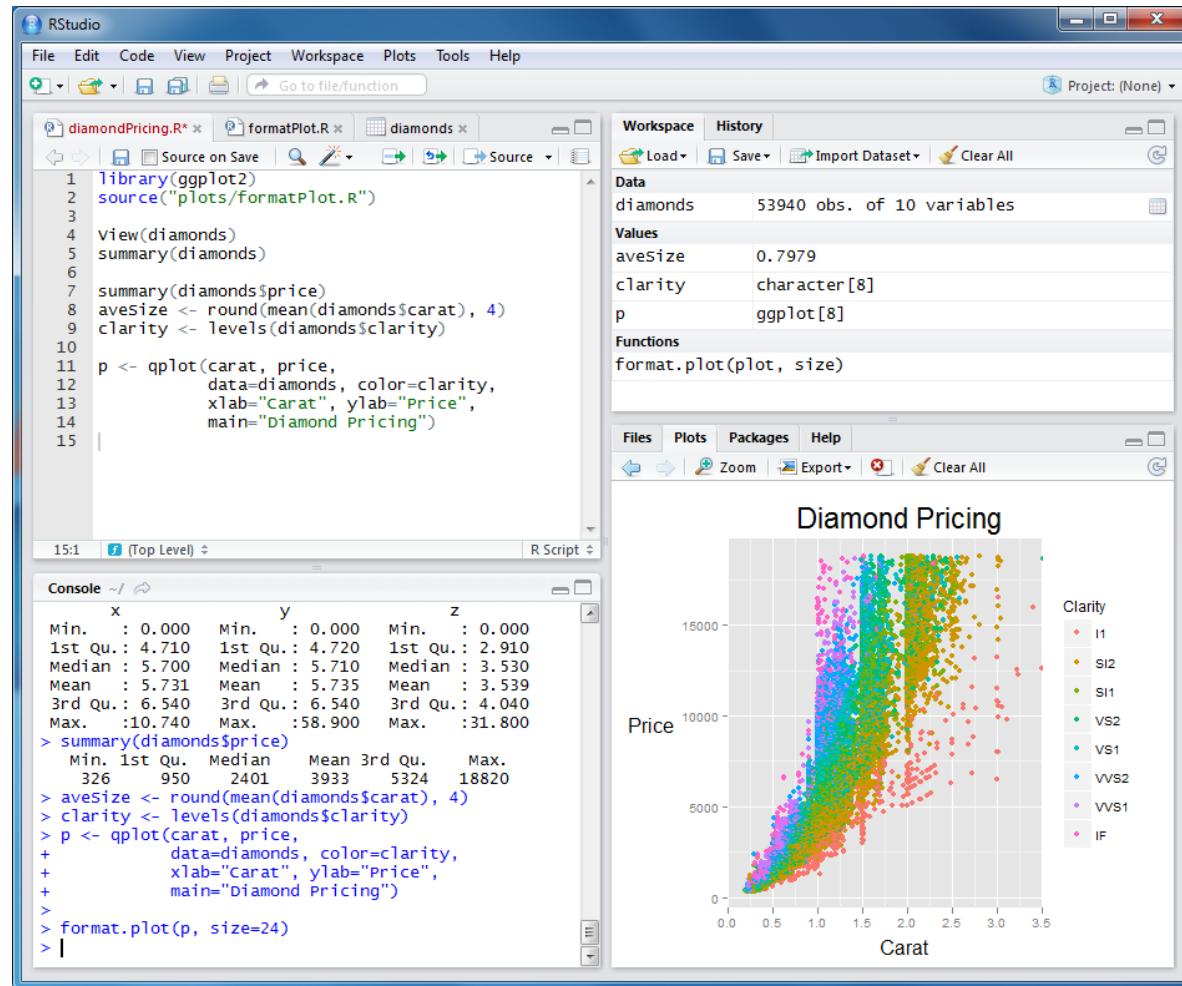
The "Support Vector Machine (LibSVM)" panel on the right provides a synopsis of the process.

Software R

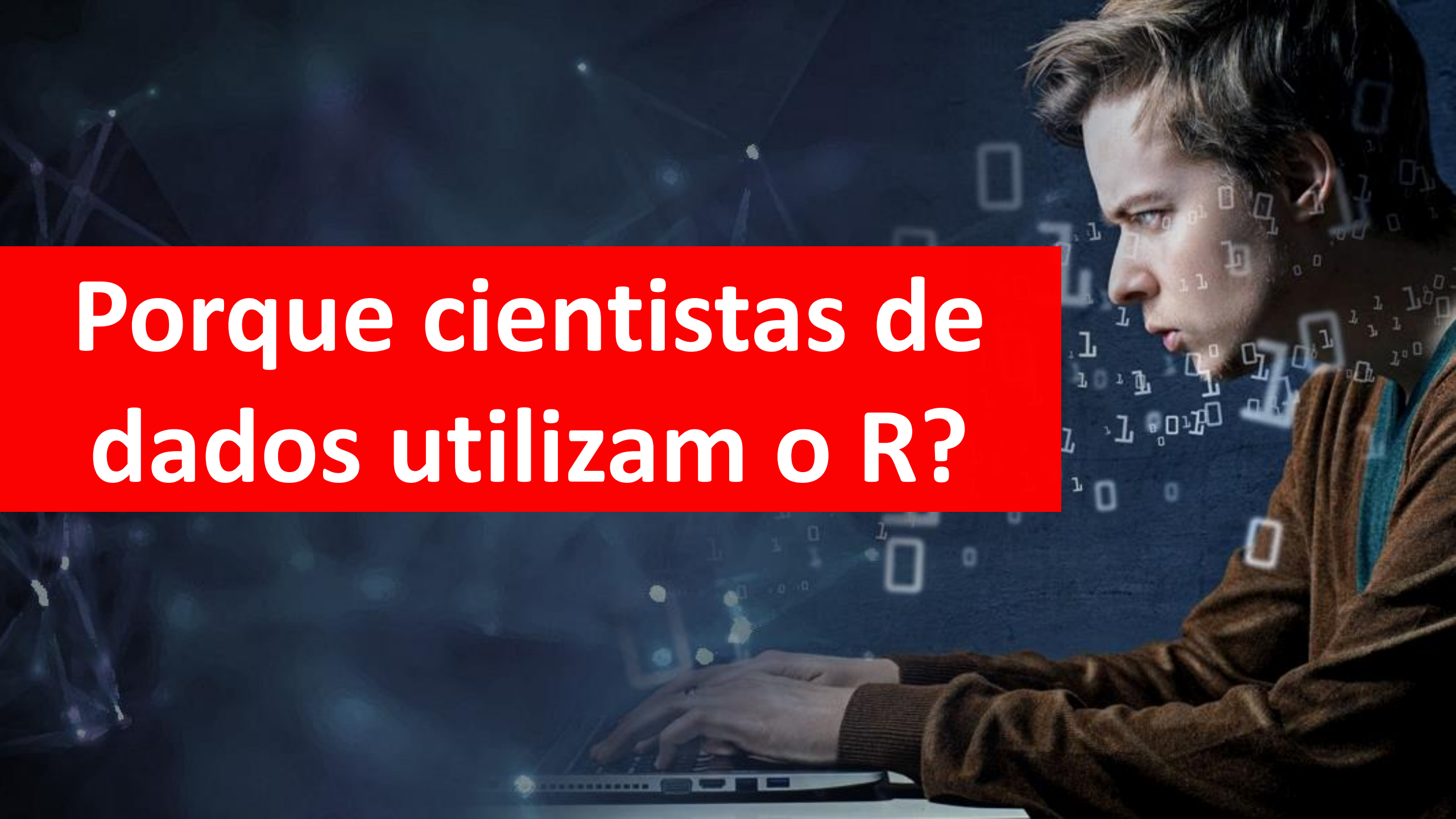
- Conhecendo softwares estatísticos;
 - ✓ É software Livre;
 - ✓ Grande quantidade de bibliotecas (pacotes);
 - ✓ É manipulado através de linha de comando;
 - ✓ Um dos mais utilizados em Data Science.



Software R



**Porque cientistas de
dados utilizam o R?**



Software Estatístico R

- Linguagem de programação especializada em computação de dados;
- É um software gratuito;
- Multiplataforma (Win, Linux, Mac...);
- Grande quantidade de bibliotecas (pacotes);



Software Estatístico R

- Foi criado por **Ross Ihaka** e **Robert Gentleman**;
- *Departamento de Estatística da universidade de Auckland, **Nova Zelândia**;*
- *O nome foi inspirado nas iniciais dos autores;*
- *Foi baseado na linguagem S (proprietária).*



Software Estatístico R

1993	Projeto de pesquisa em Auckland, na Nova Zelândia
1995	R liberado como projeto open-course
1997	Formado o grupo R-Core
2000	Liberada a versão 1.0.0 do R
2003	Criação da R Foundation
2004	Primeira conferência internacional de usuários em Vienna
2015	Formado o R Consortium (com participação da IBM e Microsoft)



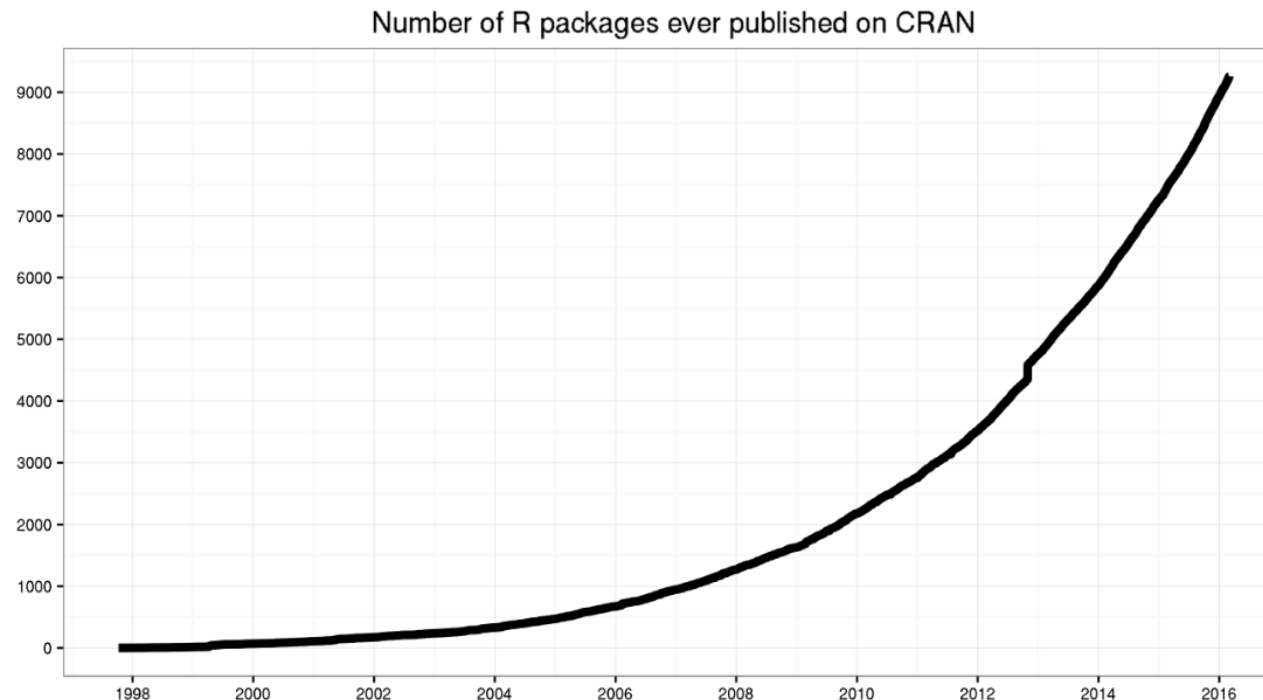
Software Estatístico R

- São disponibilizadas duas versões por ano;
- Possui funções para:
 - ✓ Extração de dados;
 - ✓ Limpeza de dados;
 - ✓ Carregamento e transformação de dados;
 - ✓ Análise estatística;
 - ✓ Machine Learning;
 - ✓ Visualização de dados;
 - ✓ ...



Software Estatístico R

- Quantidade de pacotes disponibilizados



<http://blog.revolutionanalytics.com/2016/03/16-years-of-r-history.html>

Software Estadístico R

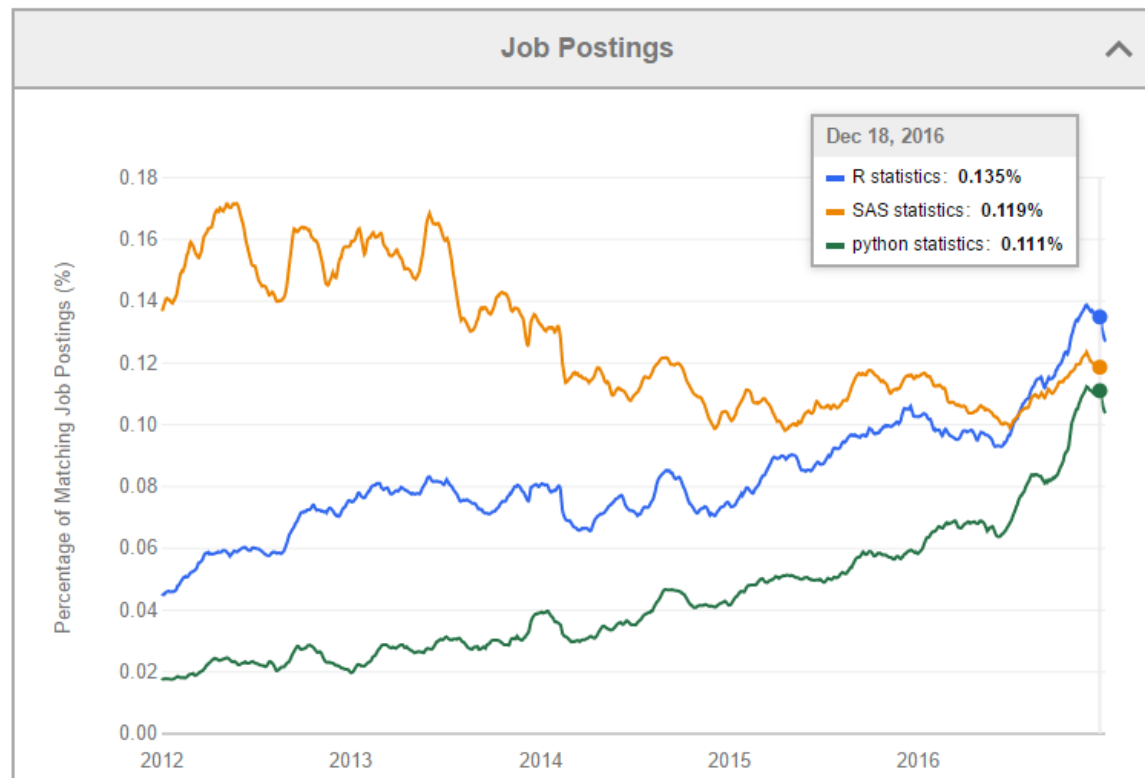
What programming languages you used for data mining / data analysis in the past 12 months? [570 voters]

R (257)	45%
SQL (184)	32%
Python (140)	25%
Java (139)	24%
SAS (121)	21%
MATLAB (83)	15%
C/C++ (73)	13%
Unix shell/awk/gawk/sed (59)	10%
Perl (45)	7.9%
Hadoop/Pig/Hive (35)	6.1%
Lisp (4)	0.7%
Other (70)	12.0%
None (7)	1.2%



Software Estatístico R

- Tendências do mercado de trabalho



<https://www.r-bloggers.com/job-trends-for-r-and-python/>



Software Estatístico R

- Vantagens e Desvantagens na utilização do R



Software Estatístico R

- Grande variedade de pacotes disponíveis gratuitamente;
- Controle total sobre o processo de análise;
- Possibilidade de integração com outras linguagens;
- Além de estatística, análises como Text Mining...



Software Estatístico R

- Grande comunidade de desenvolvedores;
- Muita documentação grátis;
- Grandes empresas investindo:



Software Estatístico R

- Grande quantidade de pacotes:
 - ✓ **sqldf** - pacote que permite realizar queries SQL em dataframes no R;
 - ✓ **forecast** - modelar séries temporais
 - ✓ **plyr** - dividir uma estrutura de dados em grupos;
 - ✓ **stringr** - manipulação de strings;
 - ✓ **database drivers** - RMongo, RODB, RMySQL;
 - ✓ **ggplot2** - visualização de dados
 - ✓ **caret** - pacote para Machine Learning;
 - ✓ quase 9.000 pacotes (<https://cran.r-project.org>).



Software Estatístico R

- Não tem uma interface gráfica robusta;
- Tudo é feito por linha de comando;
- Inicialmente tem uma curva de aprendizagem maior;



Software Estadístico R

- Sites especializados



The screenshot shows the R-bloggers website. At the top is the logo "R-bloggers" with the tagline "R news and tutorials contributed by (750) R bloggers". Below the logo is a navigation bar with links: Home, About, RSS, add your blog!, Learn R, R jobs, and Contact us. The main content area features a large article titled "Venture Capital Deals in 2016 – An Overview (2/2)" by Salvino, dated February 13, 2017. The article includes a small image of three potted plants and a "Read more" link. To the left of the main article is a sidebar with a "WELCOME!" message, a "Follow @rbloggers" button, and a "Subscribe" button. Below the sidebar is a "How to Find Equidistant Coordinates Between Two" article by Janek Thomas, dated February 12, 2017. To the right of the main article is a "RECENT POPULAR POSTS" section with a list of articles, and a "MOST VISITED ARTICLES OF THE WEEK" section with a list of articles.

WELCOME!

Follow @rbloggers 42.6K

Here you will find daily news and tutorials about R, contributed by over 573 bloggers. There are many ways to follow us - By e-mail:

Your e-mail here

Subscribe

38161 readers BY FEEDBURNER

On Facebook:

R blogg... 53 mil curtidas

Curtiu

Venture Capital Deals in 2016 – An Overview (2/2)

February 13, 2017
By Salvino



In the previous post (mostly based on EDA principles) I highlighted the main features of 1,720 Venture Capital deals that took place in 2016 in 50 different countries. It is of a central importance to once again underline that the dataset that I used is not a representative sample Read More ...

Read more »

How to Find Equidistant Coordinates Between Two

February 12, 2017
By Janek Thomas

mlr 2.10 is now on CRAN. Please update your package if you haven't done so in a while. Here is an

mlr 2.10

February 12, 2017
By Janek Thomas

mlr 2.10 is now on CRAN. Please update your package if you haven't done so in a while. Here is an

Search & Hit Enter

RECENT POPULAR POSTS

- Text mining and word cloud fundamentals in R : 5 simple steps you should know
- Dataframes and the tidyverse
- Implementing the Gradient Descent Algorithm in R

MOST VISITED ARTICLES OF THE WEEK

- How to write the first for loop in R
- Installing R packages
- Deep Learning in R
- Tutorials for learning R
- Using apply, sapply, lapply in R
- Twitter sentiment analysis with Machine Learning in R using doc2vec approach
- How to Make a Histogram with Basic R
- How to perform a Logistic Regression in R
- Scatterplots

<https://www.r-bloggers.com/>

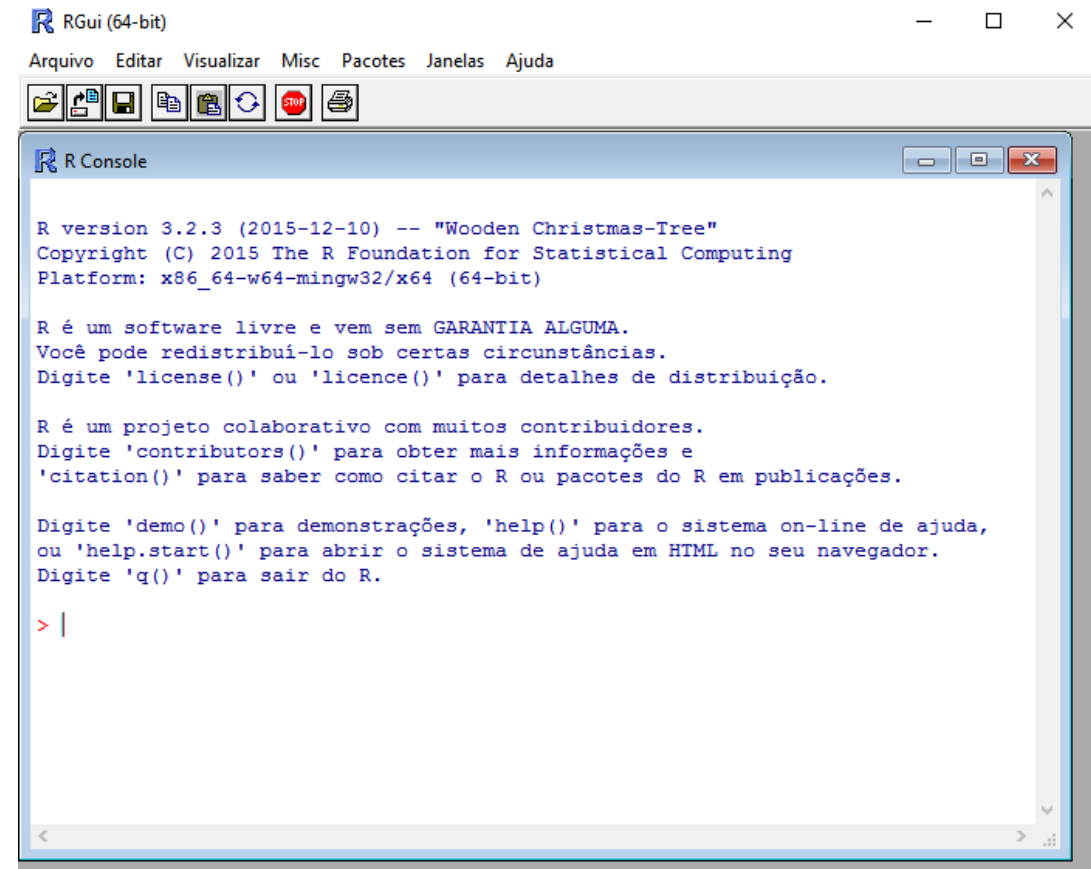


Conhecendo a Linguagem R



Interface R (Basic)

- Console básico do R;
- O console pode ser improdutivo;
- Não fornece funcionalidades para codificação;



```
RGui (64-bit)
Arquivo  Editar  Visualizar  Misc  Pacotes  Janelas  Ajuda

R Console

R version 3.2.3 (2015-12-10) -- "Wooden Christmas-Tree"
Copyright (C) 2015 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.

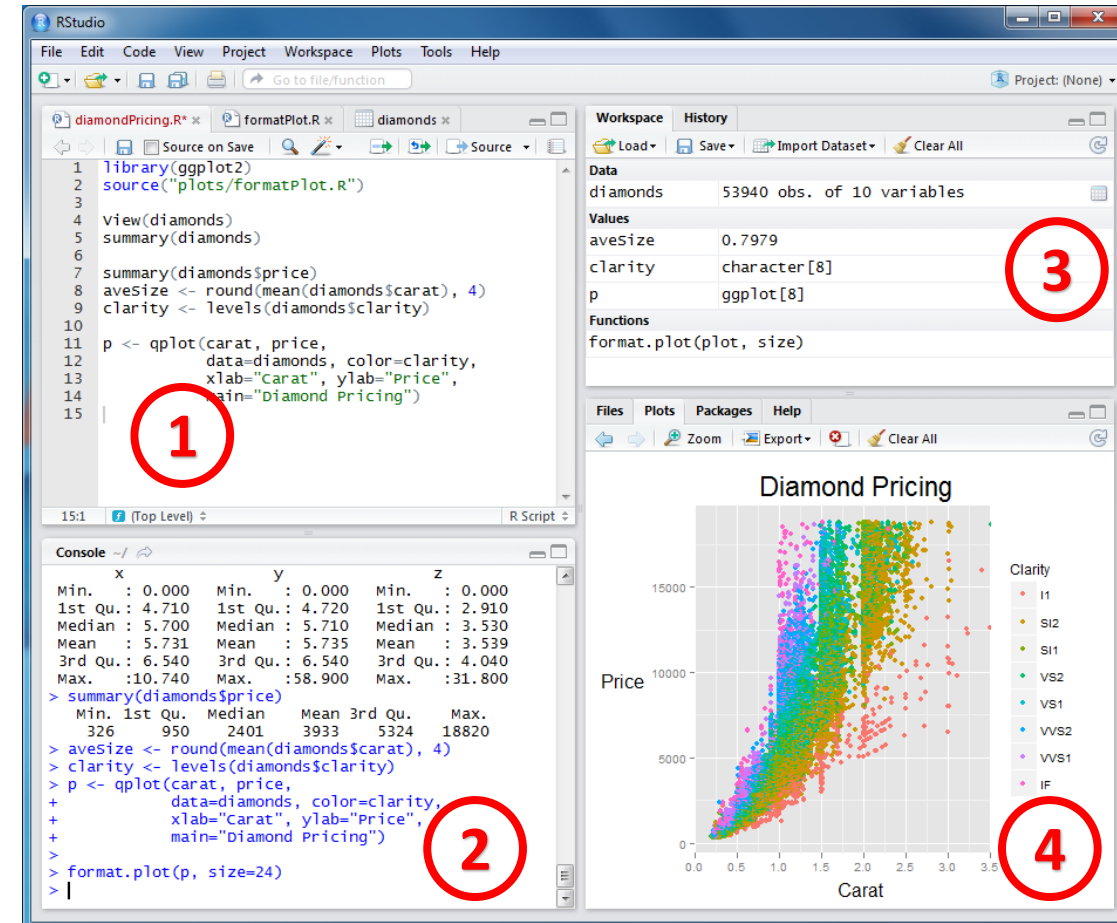
R é um projeto colaborativo com muitos contribuidores.
Digite 'contributors()' para obter mais informações e
'citation()' para saber como citar o R ou pacotes do R em publicações.

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.

> |
```

Interface R (Basic)

- **IDE – RStudio;**
 - Disponível gratuitamente;
 - Função *autocomplete*;
 - Quatro áreas básicas:
 - ✓ Codificação (1);
 - ✓ Console (2);
 - ✓ Status(3);
 - ✓ Output(4).



Conhecendo a linguagem R

- Conhecendo comandos básicos

```
1  
2 # Configurando o diretorio de trabalho  
3 getwd()  
4 setwd('C:/cursoR')
```

```
6 # Instalando pacotes no R  
7 install.packages("nomeDoPacote")  
8 library(nomeDoPacote)
```

```
12 # Documentação sobre a função  
13 help(mean)  
14 ?mean
```

Conhecendo a linguagem R

- Trabalhando com operadores

```
3 # Soma
4 4 + 4
5 # Subtracao
6 4 - 4
7 # Multiplicacao
8 4 * 4
9 # Divisao
10 4 / 4
11 # Potencia
12 4^2
13 4**2
14 # Modulo
15 14 %% 3
```

Conhecendo a linguagem R

- Operadores relacionais

```
3  a = 7
4  b = 5
5  # Operadores
6  a > 8
7  a < 8
8  a <= 8
9  a >= 8
10 a == 8
11 a != 8
12 # Operadores logicos
13 # And
14 (a==8) & (b==6)
15 # Or
16 (a==8) | (b>5)
```



Quais são os tipos de dados que o R suporta?



Tipos de dados e objetos

- Tipos de Dados em R

- ✓ Numérico;
- ✓ Character;
- ✓ Complex;
- ✓ Logic;

```
3 # Numeric
4 num = 2.5
5
6 # Character
7 char1 = 'A'
8 char1
9
10 # Complex
11 compl = 2.5 + 4i
12
13 # Logic
14 x = 1; y = 2
15 z = x > y
16 z
```

Tipos de dados e objetos

- Tipos de dados e conversões

```
3 # Numeric - Todos os numeros criados em R sao do tipo numerico
4 num = 2.5
5 num
6
7 x = -123
8 x
9 class(x)
10
11 # Conversao de tipos Numeric
12 is.integer(num)
13 y = as.integer(num)
14 y
15 class(y)
```

Tipos de dados e objetos

- Criando variáveis

```
3  # Criando Variaveis
4  var1 = 367
5  var1
6  typeof(var2)
7
8  # Criando um Vetor
9  var3 = c("seg", "ter", "qua")
10 var3
11 mode(var3)
12
13 # Criando um objeteto tipo função
14 var4 = function(x) {x+3}
15 var4(3)
16
17 # Atribuindo valores a objetos
18 x <- c(4,5,6)
19 x
20 # Verificando o valor em uma posicao especifica
21 x[1]
```

Tipos de dados e objetos

- Tipos de Objetos em R
 - O R é uma linguagem orientada a objetos;
 - Os principais tipos de objetos no R são:
 - ✓ **Vetores**: sequência de valores numéricos ou caracteres;
 - ✓ **Matrizes**: coleção de vetores em linhas e colunas;
 - ✓ **Listas**: conjuntos de vetores, matrizes e data frames;
 - ✓ **Dataframe**: mesmo que matriz mas aceita vetores com tipos de dados diferentes;
 - ✓ **Funções**: permitem os mais diversos cálculos com objetos.

Tipos de dados e objetos

- Vetores e Matrizes

```
3 # Vetor: possui 1 dimensao e 1 tipo de dado
4 vetor1 <- c(1:10)
5 vetor1
6 length(vetor1)
7 mode(vetor1)
8 typeof(vetor1)
```

```
11 # Matriz: possui 2 dimensoes e 1 tipo de dado
12 matriz1 <- matrix(1:10, nrow =2)
13 matriz1
14 length(matriz1)
15 mode(matriz1)
16 typeof(matriz1)
```


Tipos de dados e objetos

- Array e Data Frames

```
19 # Array: possui 2 ou mais dimensoes e 1 tipo de dado
20 array1 <- array(1:5, dim=c(3,3,3))
21 array1
22 length(array1)
23 mode(array1)
24 typeof(array1)
```

```
27 # Data Frames: dados de diferentes tipos
28 View(iris)
29 length(iris)
30 mode(iris)
31 typeof(iris)
```

Tipos de dados e objetos

- Listas e Funções

```
34 # Listas: colecao de diferentes objetos
35 lista1 <- list(a=matriz1, b=vetor1)
36 lista1
37 length(lista1)
38 mode(lista1)
39 typeof(lista1)
```

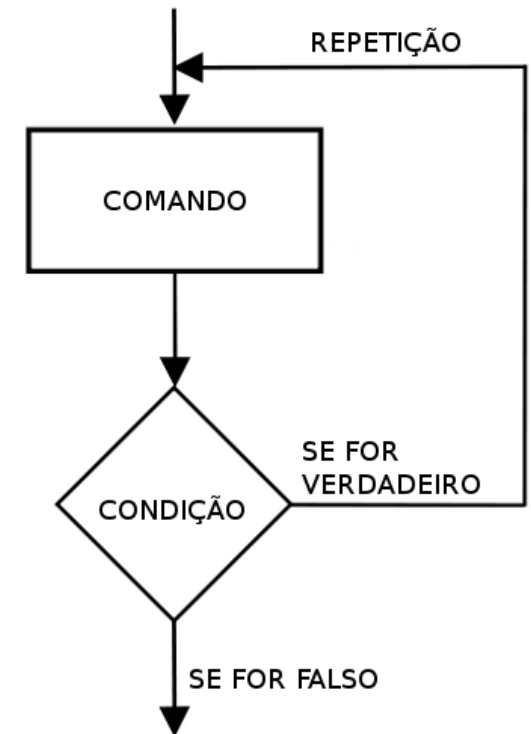
```
42 # Funções também são vistas como objetos em R
43 func1 <- function(x) {
44     var1 <- x * x
45     return(var1)
46 }
47 func1(5)
48 class(func1)
```

Estrutura de Controle



Estrutura de Controle

- Estão presentes em **todas as linguagens** de programação;
- Permite fazer **validações nos dados** e variáveis;
- Em Data Science é muito importante no **pré-processamento** dos dados.



Estrutura de Controle

- Estrutura condicional - If e Else

```
6 x = 25
7 if (x < 30)
8     {"Este numero é menor que 30"}
```

```
16 # Else
17 if (x < 7) {
18     "Este número eh menor que 7"
19 } else {
20     "Este número nao eh menor que 7"
21 }
```

Estrutura de Controle

- Estrutura condicional aninhada

```
18 # Comandos podem ser aninhados
19 x = 7
20 if (x < 7) {
21     "Este numero eh menor que 7"
22 } else if(x == 7) {
23     "Este é o numero 7"
24 }else{
25     "Este numero nao eh menor que 7"
26 }
```

Estrutura de Controle

- Funções com estruturas condicionais

```
29 # Estruturas if dentro de funcoes
30 func1 <- function(x,y){
31     ifelse(y < 7, x + y, "Não encontrado")
32 }
33
34 func1(4,2)
35 func1(40,7)
```


Estrutura de Controle

- Funções com estruturas condicionais

```
29 # Estruturas if dentro de funcoes
30 func1 <- function(x,y){
31     ifelse(y < 7, x + y, "Não encontrado")
32 }
33
34 func1(4,2)
35 func1(40,7)
```

Estrutura de Controle

- Estruturas de Loop

```
38 # Loop For
39 for (i in 1:20) {print(i)}
40 for (q in rnorm(10)) {print(q)}
41
```

```
43 # Loop While
44 x = 1
45 while(x < 5) {
46     x = x + 1
47     print(x)
48 }
```

A young man with dark hair, wearing a red and white striped shirt, is holding a magnifying glass over a chalkboard. The chalkboard is filled with various mathematical and scientific diagrams, including bar charts, line graphs, and chemical structures. The man has a focused expression, looking through the magnifying glass.

Trabalhando com funções em R

Trabalhando com funções

- Deixa o **código mais legível**, elegante e menos repetitivo;
- Funções **são objetos** em R;
- Funções podem **receber outras funções** como argumentos;
- Funções podem **chamar pacotes** específicos no R;
- É representada por:

```
1 nome_da_funcao <- function(argumentos){  
2     "corpo da funcao"  
3 }
```

Trabalhando com funções

- Funções nativas do R:

- ✓ Conhecidas como funções *Built-in*

```
6  abs ()  
7  sqrt ()  
8  mean ()  
9  c ()  
10 etc..
```

Trabalhando com funções

- Criando funções em R

```
31 # Criando funcoes
32 myfunc <- function(x) { x + x }
33 myfunc(10)
34 class(myfunc)
35
36 myfunc2 <- function(a, b) {
37     valor = a ^ b
38     print(valor)
39 }
40 myfunc2(3, 2)
41
42 jogando_dados <- function() {
43     num <- sample(1:6, size = 1) #Local
44     num
45 }
```


Trabalhando com funções

- Funções da família *Apply*

```
2  apply() - arrays e matrizes
3  tapply() - os vetores podem ser divididos em diferentes subsets
4  lapply() - vetores e listas
5  sapply() - versão amigável da lapply
6  vapply() - similar a sapply, com valor de retorno modificado
7  rapply() - similar a lapply()
8  eapply() - gera uma lista
9  mapply() - similar a sapply, multivariada
```

Trabalhando com funções

- Exemplo: funções *Apply*

```
# Usando um Loop  
lista1 <- list(a = (1:10), b = (45:77))
```

```
# Calculando o total de cada lista  
valor_a = 0  
valor_b = 0  
for (i in lista1$a){  
  valor_a = valor_a + i  
}  
for (j in lista1$b){  
  valor_b = valor_b + j  
}  
print(valor_a)  
print(valor_b)
```

VS

```
# Calculando o total de cada lista  
?apply  
apply(lista1, sum)
```

```
> apply(lista1, sum)  
      a      b  
55 2013
```

```
> print(valor_a)  
[1] 55  
> print(valor_b)  
[1] 2013
```

Dúvidas



- **Contatos:**

- ✓ Email: rodrigo.linsrodrigues@ufrpe.br
- ✓ Facebook: [/rodrigomuribec](https://www.facebook.com/rodrigomuribec)