



THE
DEVELOPER'S
CONFERENCE

Trilha – BIG DATA

Data Analysis com R

Rodrigo Ribeiro Gonçalves



Rodrigo Ribeiro Gonçalves



THE
DEVELOPER'S
CONFERENCE

- **Atualmente**
 - Administrador de Banco de Dados
 - Professor Universitário
- Email: dbconsultoria@gmail.com

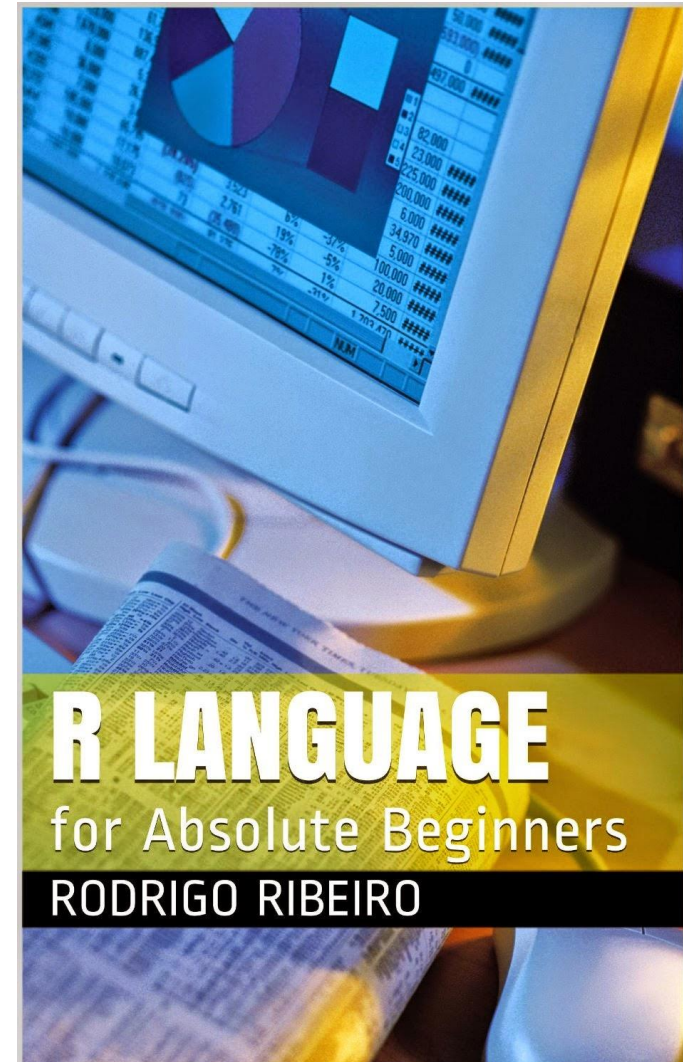


Rodrigo Ribeiro Gonçalves



THE
DEVELOPER'S
CONFERENCE

- Autor do livro [Linguagem R – Para Iniciantes](#)
- Autor de outros livros e e-books sobre tecnologia da informação.
- Blog:
www.tsqldata.blogspot.com



➤ **Parte Teórica**

- BIG DATA e Mercado
 - Problema ou solução
- A Linguagem R
 - O que é e como usar

➤ **Parte Prática**

- Usando o R
 - Obtendo dados
 - Analisando Dados
 - Apresentando Resultados

➤ **Encerramento**

- Outros recursos
- Considerações



THE
DEVELOPER'S
CONFERENCE



THE
DEVELOPER'S
CONFERENCE

Segundo dados do Gartner Research (2014), mais de 50% dos projetos de Big Data e Analytics **falham**.

- Estouro de orçamento
- Estouro de prazo
- Má definição de escopo
- Má qualidade da entrega

Fonte: <http://www.analytics-magazine.org/july-august-2014/1074-the-data-economy-why-do-so-many-analytics-projects-fail>





THE
DEVELOPER'S
CONFERENCE

Falta de apoio da parte estratégica da empresa.



Por quê?



THE
DEVELOPER'S
CONFERENCE

- Analytics envolve olhar para o **passado**... rever desenvolvimentos já realizados (resistência).
- Para acessar os dados é necessário quebrar/dobrar as barreiras de segurança da informação pré-estabelecidas.
- Interagir com pessoas que estão envolvidas em outros projetos ou que não tem interesse em participar nos projetos de Analytics.

Desafios



THE
DEVELOPER'S
CONFERENCE

- Descolamento entre expectativa e realidade, devido a pouca cultura de Analytics (Analfabetismo de Dados).
- Um pouco de confusão em relação ao termo BIG DATA e os resultados que podem ser obtidos através de um projeto de BIG DATA e BIG DATA ANALYTICS.

BIG DATA, ANALYTICS e DATA SCIENCE



THE
DEVELOPER'S
CONFERENCE

- BIG DATA e DATA ANALYTICS **NÃO** é um produto ou um computador
- BIG DATA e DATA ANALYTICS é uma **estratégia, visão e arquitetura** em constante evolução.
- BIG DATA e DATA ANALYTICS **envolve pessoas** (Cientistas de dados, Analista de Dados, Analista de Requisitos, DBAs, Infra-estrutura, Investimento em treinamentos, capacitação, software, etc...)

Linguagem R



THE
DEVELOPER'S
CONFERENCE

**Por ser uma tecnologia voltada para estatística,
se encaixa bem no nicho de DATA SCIENCE.**

- **Data Analysis**
- **Machine Learning**
- **Text Mining**
- **Estatística**
- **Data Visualization**
 - **Gráficos, histogramas... etc**

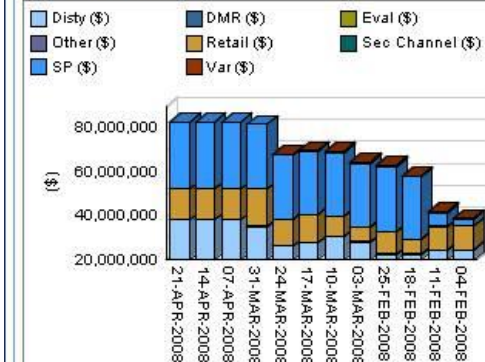
R pode ser usado
para BI?



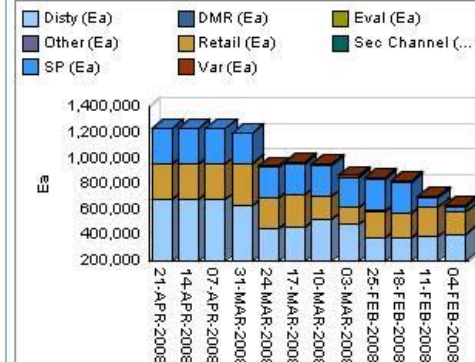
THE
DEVELOPER'S
CONFERENCE

- **Business Intelligence usa estatística descritiva.**
 - Foco na visualização e **SUMARIZAÇÃO!**
- **BIG DATA e DATA SCIENCE usam estatística dedutiva para analisar e realizar previsões de resultados e de comportamentos**
 - Foco nos **INSIGHTS!**

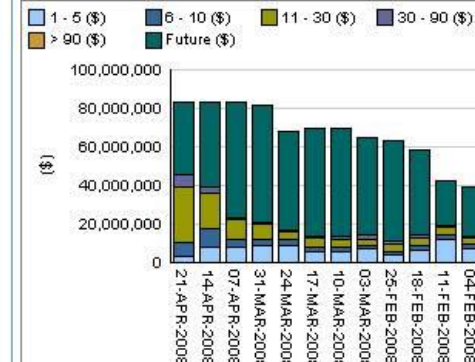
Backorder Amount (\$)



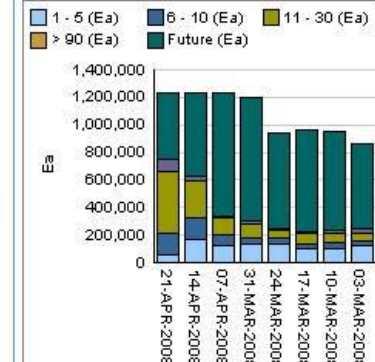
Backorder Qty



Backlog Aging (\$)



Backlog Aging Units



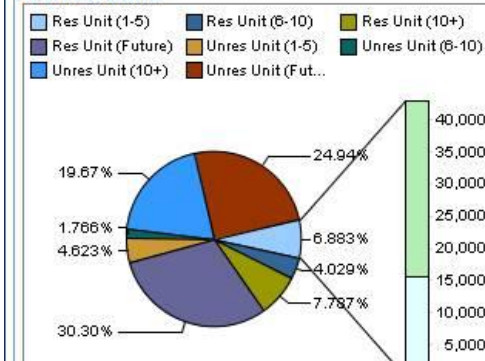
Personalize Table Layout

New

Personalize Table Layout

Edit Select

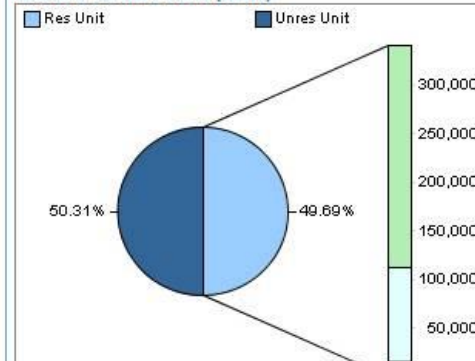
Unit Distribution



Personalize Table Layout

Edit Select

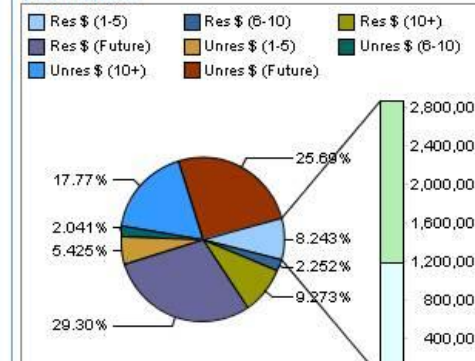
Reserve Vs Unreserve (Units)



Personalize Table Layout

Edit Select

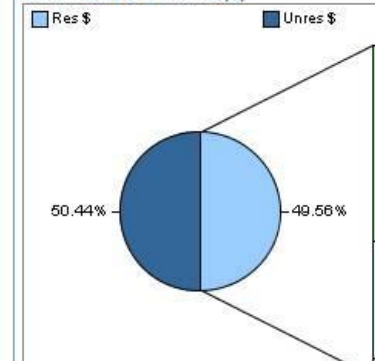
\$ Distribution



Personalize Table Layout

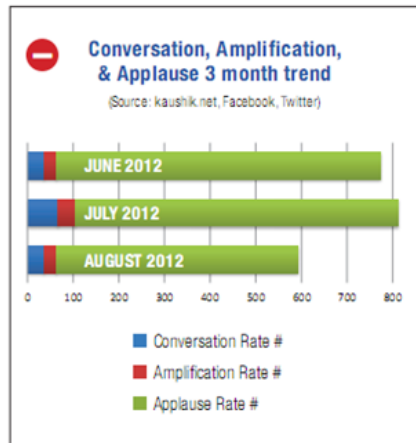
Edit Select

Reserve Vs Unreserve (\$)



- Típico dashboard de BI!

Behavior Strategy: Increase visitor engagement and loyalty



Dear Avinash: Attribution Modeling, Org Culture, Deeper Analysis +++
Published: August 13, 2012

Conversation Rate #
36 Comments

Amplification Rate #
Retweets: 25
Facebook shares: 3

Applause Rate #
Stumbleupon: 1
Facebook: 78
Twitter: 329
Google+: 66
LinkedIn: 55

Visits to post: 5,955 (30 days after post)

Visits that spent 3 or more minutes on post: 737

Visits that spent 3 or more minutes on post and viewed 3 or more pages: 337

Frequency - Count of visits +9: 11,740
% of Total: 9.36% (125,484)

Frequency - < 5 days Since Last Visit: 113,520
% of Total: 90.47% (125,484)



Web Analytics Consulting For Smarter Decisions
Published: July 23, 2012

Conversation Rate #
67 Comments

Amplification Rate #
Retweets: 33
Facebook shares: 5

Visits to post: 2,404 (30 days after post)

Visits that spent 3 or more minutes on post: 737

Visits that spent 3 or more minutes on post and viewed 3 or more pages: 337

Frequency - Count of visits +9: 11,740
% of Total: 9.60% (136,073)

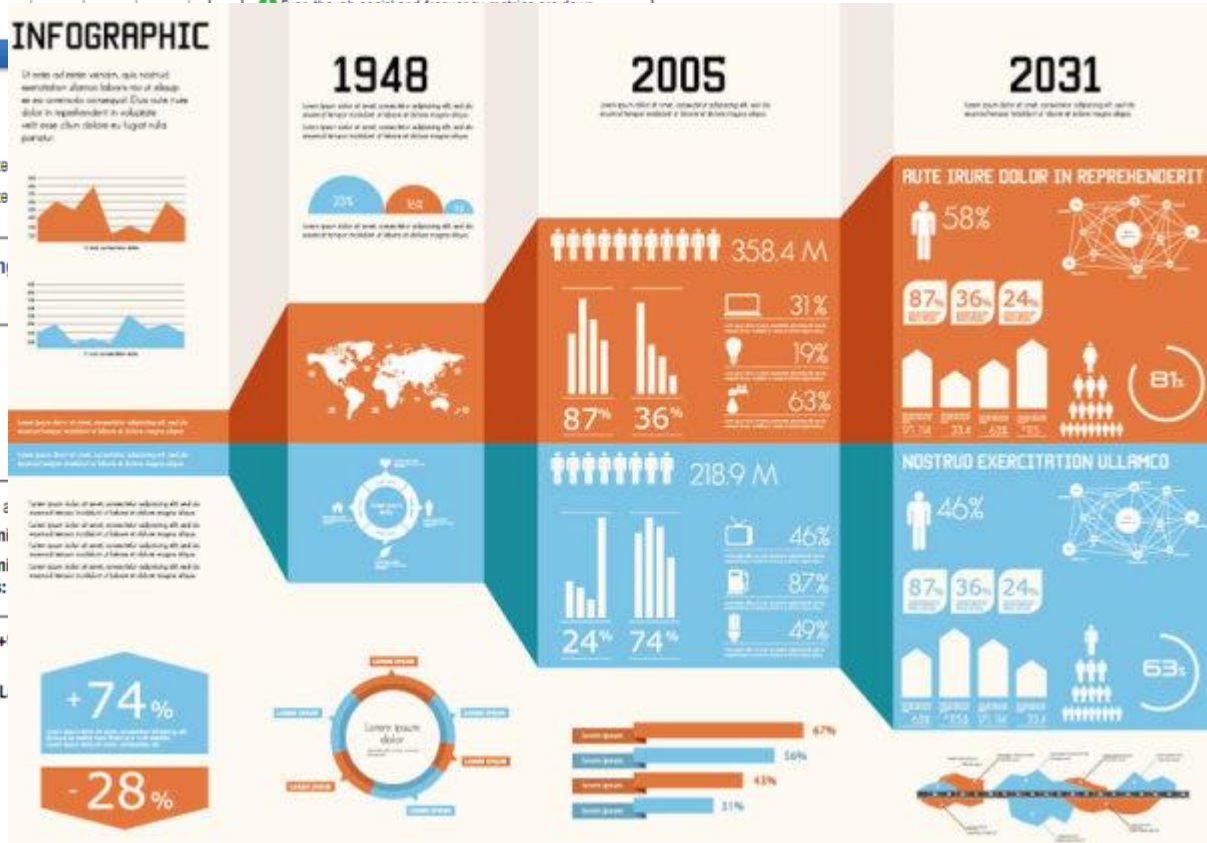
Frequency - < 5 days Since Last Visit: 113,520
% of Total: 89.74% (136,073)

Key Trends & Insights:

- In Aug 2012, Conversation Rate is down by 47%, Amplification Rate is down by 27%, and Applause Rate is down 26% from the previous month.
- Frequency Rates are also down. Visitors who visited more than 9 times in the month is down 11%, and visitors who returned in less than 5 days is down 8%.



THE DEVELOPER'S CONFERENCE



- Relatório de Data Science: Não apenas dados e gráficos, mas também INSIGHTS!**

Linguagem R



THE
DEVELOPER'S
CONFERENCE

- É uma linguagem e um ambiente de desenvolvimento.
- R é uma evolução da linguagem S, desenvolvida dentro do BELL LABS.
- É um software FREE e OPEN SOURCE, distribuído pela GPL Versão 2
- Pode ser obtido no repositório do projeto: <http://cran.r-project.org> (Comprehensive R Archive Network)
- Lançada publicamente em 1993
- Criado para trabalhar com grandes DATASETS

Revolution R OPEN

- Uma distribuição alternativa do R com *features* adicionais, além de total compatibilidade com o R “original”.
- A empresa Revolution Analytics acabou de ser comprada pela Microsoft! [Fonte](#)

Linguagem R



THE
DEVELOPER'S
CONFERENCE

- O R pode ser instalado no Windows, Linux e Mac com versões 32 e 64 bits
- Depois de instalado, pode ser usado através de uma linha de comando ou através de uma GUI instalada junto com o ambiente de desenvolvimento

```
C:\Windows\system32\CMD.exe  
am Files\R\R-3.1.2\bin>R.exe  
n 3.1.2 (2014-10-31) -- "Pumpkin Helmet"  
t (C) 2014 The R Foundation for Statistical Computing  
: i386-w64-mingw32/i386 (32-bit)
```

software livre e vem sem GARANTIA ALGUMA.
e redistribuí-lo sob certas circunstâncias
license()' ou 'licence()' para detalhes de
projeto colaborativo com muitos contribuidores
contributors()' para obter mais informações
n()' para saber como citar o R ou pacotes
demo()' para demonstrações, 'help()' para
.start()' para abrir o sistema de ajuda em
q()' para sair do R.

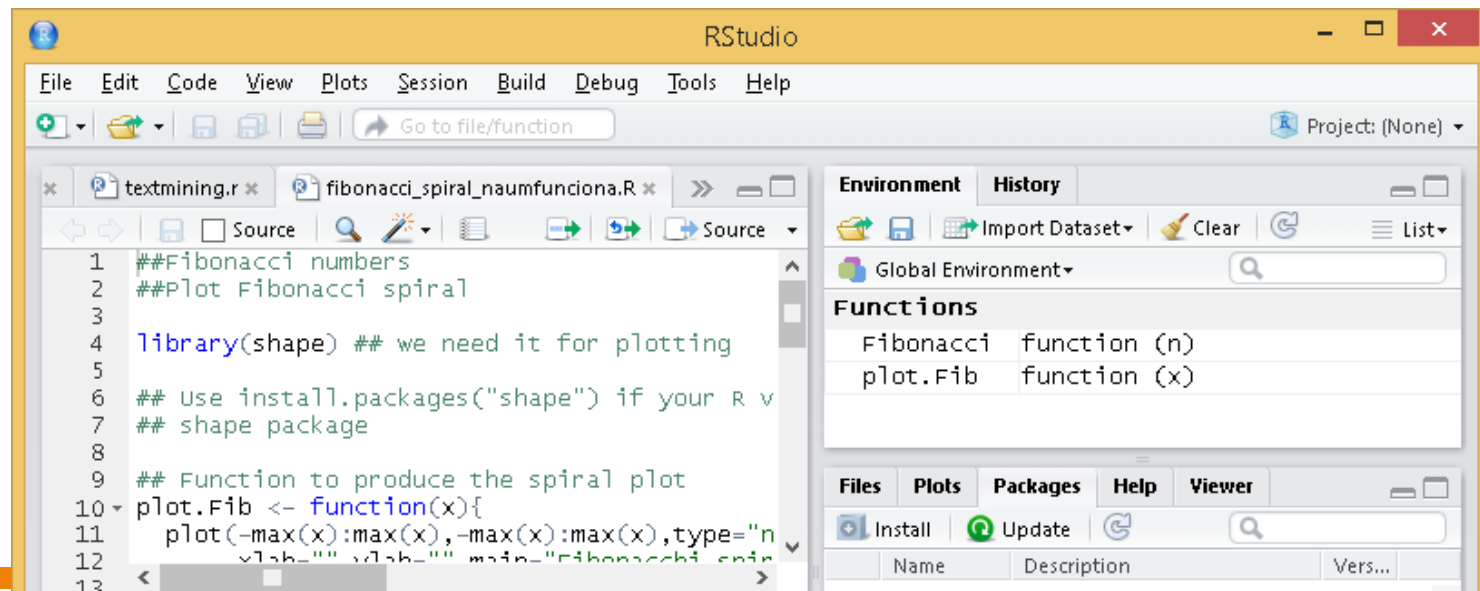
```
RGui (64-bit) - [R Console]  
Arquivo  Editar  Visualizar  Misc  Pacotes  Janelas  Ajuda  
R version 3.1.2 (2014-10-31) -- "Pumpkin Helmet"  
Copyright (C) 2014 The R Foundation for Statistical Computing  
Platform: x86_64-w64-mingw32/x64 (64-bit)
```


R STUDIO?



THE
DEVELOPER'S
CONFERENCE

- O R também pode ser usado através de uma interface também FREE, chamada de RSTUDIO distribuído pela AGPL V3
- WWW.RSTUDIO.COM



Extensões



THE
DEVELOPER'S
CONFERENCE

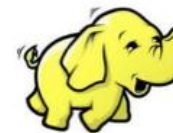
- **É uma linguagem bastante extensível através de pacotes desenvolvidos pela comunidade, vantagem que outras tecnologias não tem.**
- <http://cran.r-project.org>

Linguagem R



THE
DEVELOPER'S
CONFERENCE

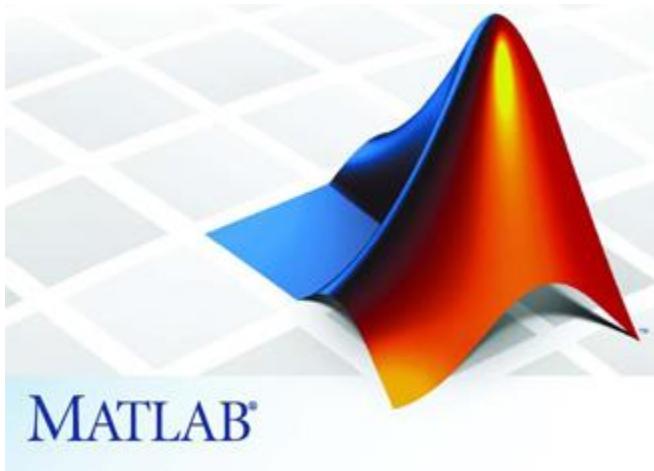
- Linguagem de SCRIPT interpretada
- Linguagem Orientada a Objetos
- Pode incorporar com relativa facilidade, dados de diferentes origens, como arquivo txt, excel, web, banco de dados relacionais, NoSQL, etc...



Concorrentes



THE
DEVELOPER'S
CONFERENCE



- **Matlab: U\$ 2.650,00**
- **STATA: U\$ 1.195,00**
- **Não apenas o licenciamento é diferente...**

Linguagem R



THE
DEVELOPER'S
CONFERENCE



- **Limitação com volumes de dados (até 1 milhão de linhas)**



THE
DEVELOPER'S
CONFERENCE

<https://github.com/dbconsultoria/tdc2015>

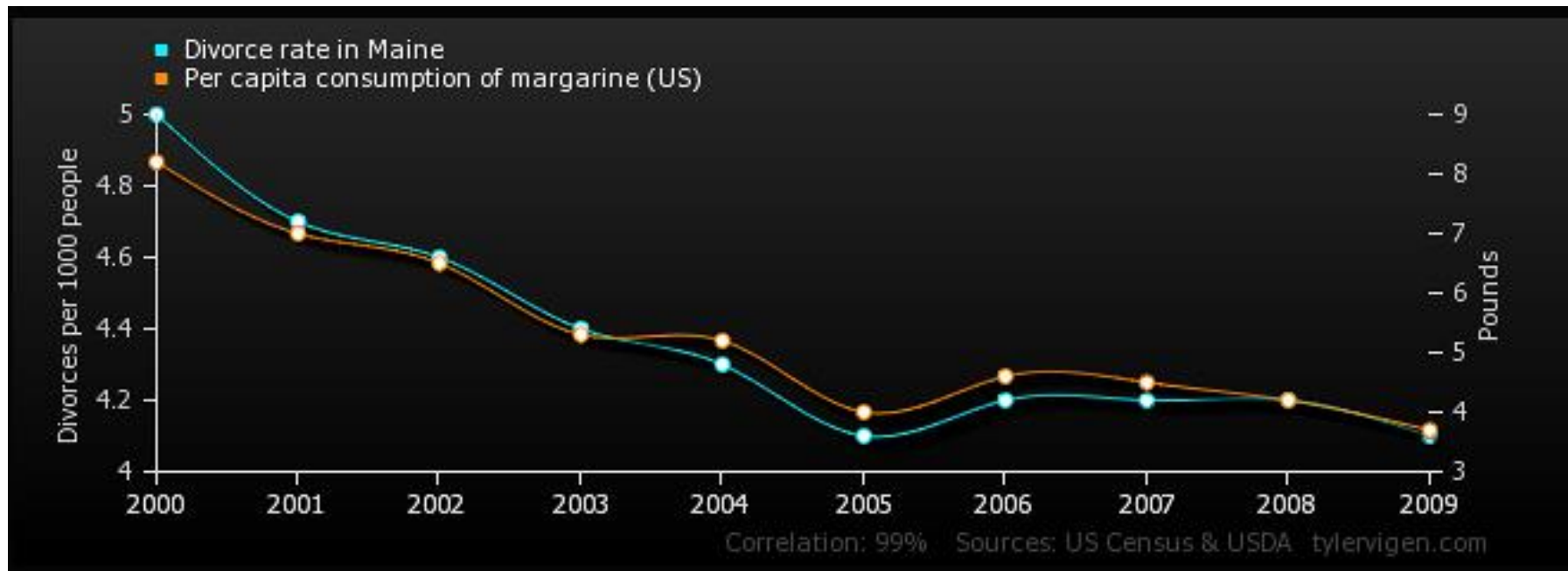


Linguagem R



THE
DEVELOPER'S
CONFERENCE

- “Correlation doesn't imply causation”



Exemplo 1



THE
DEVELOPER'S
CONFERENCE

Atribuições e operações básicas.

Exemplo 2



THE
DEVELOPER'S
CONFERENCE

Vetores e funções úteis.

Exemplo 3



THE
DEVELOPER'S
CONFERENCE

`Data frames.`

Exemplo 4



THE
DEVELOPER'S
CONFERENCE

Importando arquivos CSV.

Exemplo 5



THE
DEVELOPER'S
CONFERENCE

Importando arquivos CSV da
internet.

Exemplo 6



THE
DEVELOPER'S
CONFERENCE

**Instalando pacotes (RMySQL) e
conectando com o MySQL.**

Exemplo 7



THE
DEVELOPER'S
CONFERENCE

Gráficos básicos usando o plot.

Exemplo 8



THE
DEVELOPER'S
CONFERENCE

**Gráficos de linha com o Dataset
Oranges.**

Exemplo 9



THE
DEVELOPER'S
CONFERENCE

**Histogramas e curva normal com o
Dataset MTCARS.**

Exemplo 10



THE
DEVELOPER'S
CONFERENCE

Dotplots com o Dataset MTCARS.

Exemplo 11



THE
DEVELOPER'S
CONFERENCE

PieChart com dados da UCLA.

Outros Recursos

- **R-Bloggers:** www.r-bloggers.com
- **StackOverflow:**
<http://stackoverflow.com/questions/tagged/r>
- **Listas:** r-br@listas.c3sl.ufpr.br



THE
DEVELOPER'S
CONFERENCE

coursera
education for everyone

Perguntas e Dúvidas



THE
DEVELOPER'S
CONFERENCE

