Vítor Will

Sobre o Auto

O mundo dos dados

Getting and Clear

Exploratory Dat

Modeling

Communicatio

O mundo do F

.. ...

Contate



R e *Data Science*: como entender o que dizem os dados com o **R**?

Palestra na HackTown 2018 - INATEL

Vítor Wilher

analisemacro.com.br

13 de Dezembro de 2019

Vítor Wilh

Sobre o Auto

O mundo dos dados

Getting and Cleanin, data

Exploratory Data Analysis

Modeling Communication

O mundo do F

Um problema simples

Contato

O plano de voo para hoje

- Sobre o Autor
- O mundo dos dados
 - Getting and Cleaning data
 - Exploratory Data Analysis
 - Modeling
 - Communication
- O mundo do R
- 4 Um problema simples
- Contato

Vítor Wilh

Sobre o Autor

O mundo dos

Getting and Clear data

Exploratory Data Analysis

Modeling Communication

O mundo do l

Um problema simples

Contato

Sobre o Autor

Vítor Wilher é Bacharel e Mestre em Economia, pela Universidade Federal Fluminense, com especialização em Data Science pela Johns Hopkins University. Sócio-Fundador da Análise Macro, empresa especializada em treinamento e consultoria em data science. É também Conselheiro do Instituto Millenium.

Maiores informações, visite www.analisemacro.com.br

Vítor Wilh

Sobre o Auto

O mundo dos dados

Getting and Cleanii data

Exploratory Data

Modeling Communication

O mundo do F

Um problema simples

Contato

O mundo dos dados

O avanço da informática e das telecomunicações possibilitou o armazenamento e a distribuição de conjuntos de dados cada vez mais complexos. Lidar com essas bases de dados exigiu a sistematização de diversas técnicas de coleta, tratamento, análise e apresentação de dados.

Vítor Wilh

Sobre o Autor

O mundo dos dados

data
Exploratory Data
Analysis
Modeling
Communication

O mundo do R

Um problema simples

O mundo dos dados

Essa sistematização de técnicas deu origem ao que hoje chamamos de **data science**, cujo objetivo principal é extrair informações úteis de conjuntos de dados aparentemente confusos.

Aplicações interessantes:

- Identificar mensagens indesejáveis em um e-mail (spam);
- Segmentação do comportamento de consumidores para propagandas direcionadas;
- Redução de fraudes em transações de cartão de crédito;
- Predição de eleições;
- Otimização do uso de energia em casas ou prédios;
- etc, etc, etc...

Vítor Wilh

Sobre o Autor

O mundo dos dados

Exploratory Data Analysis Modeling

O mundo do F

Um problema

Contato

O mundo dos dados

De modo a responder esse tipo de pergunta, é necessário cumprir aquelas quatro etapas da ciência de dados.

As quatro operações:

- É preciso coletar os dados;
- Dados brutos precisam ser tratados;
- Uma vez disponíveis, os dados precisam ser analisados de forma a extrair informações relevantes e/ou responder determinados questionamentos;
- Com as respostas em mãos, é preciso apresentar os resultados.

Vítor Wilh

Sobre o Autor

O mundo dos dados

> Getting and Clear data

Analysis Modeling

Communication

llm problems

Um problema simples

Contato

O mundo dos dados

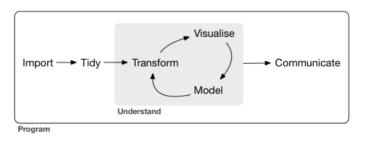


Figura: Fonte: R for Data Science.

Cada uma dessas etapas exige conhecimentos específicos, de modo a lidar com diferentes formatos de dados, bem como responder questões distintas.

Getting and Cleaning

Getting and Cleaning data

Dados podem estar dispostos em diferentes formatos:

- Excel:
- XML:
- JSON:
- txt:
- HTML:
- MySQL;
- Formatos proprietários (Weka, Stata, Minitab, Octave, SPSS, SAS, etc).

Vítor Wilh

Sobre o Auto

O mundo dos dados Getting and Cleaning

Exploratory Data Analysis

Analysis Modeling Communication

O mundo do l

-

Getting and Cleaning data

Dados precisam ser tratados:

- Limpeza de dados;
- Tratamento de missing values;
- Construção de números índices;
- Deflacionar valores correntes;
- Obtenção de taxas de crescimento, a partir de comparações mensais, interanuais, acumuladas em 12 meses, etc;
- Tratando tendências:
- Dessazonalização;
- Obtendo subconjuntos (subsetting) relevantes;
- Classificando dados de acordo com algum critério;
- Transformando dados de acordo com alguma operação.

Vítor Wilh

Sobre o Auto

O mundo do: dados

Getting and Cl

Exploratory Data Analysis

Communication

O mundo do R

Um problema simples

Contato

Exploratory Data Analysis

Dados precisam ser visualizados:

- Gráficos simples;
- Gráficos de correlação;
- Clustering;

Vítor Will

Sobre o Auto

O mundo dos dados

data

Exploratory Data Δηρίωσε

Modeling

O mundo do I

O manao do n

Um problema

Contato

Modeling

Dados podem ser relacionados uns aos outros.

- Modelos ARIMA;
- Regressão linear;
- Árvores de regressão;
- Neural Network;
- Support Vector Machine;
- Naive Bayes;
- etc, etc, etc.

litor Wilh

Sobre o Auto

O mundo dos dados

Getting and Cleani

Exploratory Da Analysis

Communication

O mundo do R

Um problema simples

Contato

Communication

Os resultados precisam ser comunicados através de *documentos reprodutíveis*, que unam **código** e **texto**.

Vítor Wilh

Sobre o Autor

O mundo do: dados

data Exploratory Data Analysis Modeling

O mundo do R

Jm problema

C--+-+-

O mundo do R

Era necessário construir uma plataforma que unisse todas essas etapas. O $\bf R$ é uma das melhores soluções atualmente disponíveis, dados os seguintes motivos:

- A existência de uma comunidade grande e bastante entusiasmada, que compartilha conhecimento todo o tempo;
- o R é gratuito, open source, de modo que você não precisa comprar licenças de software para instalá-lo;
- Tem inúmeras bibliotecas (pacotes) em estatística, machine learning, visualização, importação e tratamento de dados;
- Possui uma linguagem estabelecida para data analysis;
- Ferramentas poderosas para comunicação dos resultados da sua pesquisa, seja em forma de um website ou em pdf.

Vítor Wilh

Sobre o Auto

O mundo dos

Getting and C

Exploratory Data Analysis Modeling

O mundo do R

Um problema simples

Contato

O mundo do R

Ao aprender **R**, você conseguirá integrar as etapas de coleta, tratamento, análise e apresentação de dados em um único ambiente. Você vai esquecer ter de abrir o excel, algum pacote estatístico, depois o power point ou o word, depois um compilador de pdf para gerar seu relatório. Todas essas etapas serão feitas em um único ambiente. E essa talvez seja a grande motivação para você entrar de cabeça nesse mundo.

Vítor Wilh

Sobre o Auto

O mundo dos

Getting and Cleanii data Exploratory Data Analysis

Modeling Communication

O mundo do R

Um problema simples

Contato

O mundo do R

- Baixe o R em http://cran-r.c3sl.ufpr.br/;
- Baixe o RStudio em https://www.rstudio.com/products/rstudio/download/;
- Baixe o MikTex se você for usuário de Windows em http://miktex.org/download;
- Baixe o MacTex se você for usuário de Mac em http://www.tug.org/mactex/.

..

Sabra a Autor

O mundo dos

Getting and Cle

data Exploratory Data Analysis

O mundo do R

Um problema

C--+-+.

O mundo do R

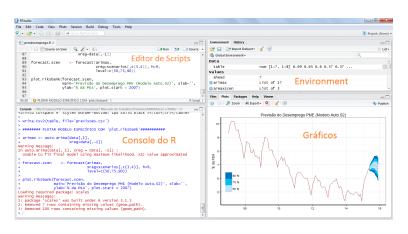


Figura: Ambiente do RStudio.

Vítor Will

Sobre o Auto

O mundo dos dados

Getting and Clean data Exploratory Data

Modeling Communication

O mundo do F

Um problema

simples

Um problema simples

Para terminar, apenas um exemplo simples de como podemos nos beneficiar das técnicas de data science para explorar diversos problemas. No ano passado, houve intensa discussão sobre aplicativos de transporte no país, dado o trâmite do PLC 28/2017 no Congresso Nacional, que buscou regulamentar a atividade. Inspirados por essa controvérsia, podemos querer entender se existe uma relação de causalidade entre procuras pela Uber e a taxa de desemprego. A hipótese implícita nesse estudo é a de que o aumento recente do desemprego teve influência no número de motoristas cadastrados na Uber e em outros aplicativos de transporte.

Vítor Will

Sobre o Auto

O mundo dos dados Getting and Cleanii data Exploratory Data Analysis

O mundo do R

Um problema

simples

Um problema simples

De modo a analisar essa questão, alguns problemas imediatos surgem:

- Onde estão os dados?
- Qual proxy utilizar para representar o interesse pela Uber?
- 3 Como tratar os dados brutos obtidos das fontes primárias?
- Qual a estrutura dos dados?
- Uma vez que as questões anteriores estejam resolvidas, qual o melhor modelo para analisar a relação entre as variáveis?

Vítor Wilh

Sobre o Auto

O mundo dos

Getting and Clean

Exploratory Da Analysis

Modeling

^ . .

Um problema

simples

Um problema simples

Vamos abrir o RStudio e começar a brincar?

/ítor Wilh

Sobre o Auto

O mundo do: dados

detting and Cit

Exploratory Data Analysis Modeling

Communication

Um problema

Contato

Slides estão disponíveis no repositório da Análise Macro no Github: https://github.com/analisemacro/degustacao.

Visite:

www.analisemacro.com.br

