



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

IMIL na Sala de Aula - IE/UFRJ

Vítor Wilher

analisemacro.com.br

9 de Novembro de 2018

O plano de voo para hoje

- 1 Sobre o Autor
- 2 O mundo dos dados
 - Getting and Cleaning data
 - Exploratory Data Analysis
 - Modeling
 - Communication
- 3 O mundo do R
- 4 Alguns problemas práticos
- 5 Data Science e os Economistas
- 6 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

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.

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...

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.

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).

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.

Exploratory Data Analysis

Dados precisam ser visualizados:

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

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.

Communication

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

O mundo do R

Era necessário construir uma plataforma que unisse todas essas etapas. O **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.

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.

O mundo do R

- 1 Baixe o **R** em <http://cran-r.c3sl.ufpr.br/>;
- 2 Baixe o **RStudio** em <https://www.rstudio.com/products/rstudio/download/>;
- 3 Baixe o **MikTeX** *se você for usuário de Windows* em <http://miktex.org/download>;
- 4 Baixe o **MacTeX** *se você for usuário de Mac* em <http://www.tug.org/mactex/>.

Alguns problemas práticos

- 1 Existe relação entre desemprego e procura pela Uber?
- 2 Os desembolsos do BNDES fizeram aumentar a taxa de investimento da economia brasileira?
- 3 Como podemos explicar a inflação de alimentos?
- 4 O Banco Central brasileiro reage a choques cambiais?
- 5 Qual o efeito do aumento de volatilidade no mercado sobre a taxa de câmbio?

Alguns problemas práticos

De modo a analisar essas questões, alguns problemas imediatos surgem:

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

Data Science e os Economistas

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

Vítor Wilher

Sobre o Autor

O mundo dos dados

Getting and Cleaning data

Exploratory Data Analysis

Modeling

Communication

O mundo do R

Alguns problemas práticos

Data Science e os Economistas

Contato

Os economistas estão acostumados com **modelagem** de variáveis econômicas, de modo que podem e devem se beneficiar das técnicas de *data science*, que envolvem coleta, tratamento, análise e apresentação de dados.

Em um mundo onde se discute a perda de empregos para robôs, os economistas devem se preparar para temas como *Big Data*, *Machine Learning*, *Statistical Learning*, *Predictive Modeling*, etc...

Data Science e os Economistas

Para terminar, uma provocação. Enquanto vocês são obrigados a ler Marx, no MIT...

Computer Science, Economics, and Data Science (Course 6-14)

Computer Science, Economics, and Data Science

Bachelor of Science in Computer Science, Economics, and Data Science

General Institute Requirements (GIRs)

The General Institute Requirements include a Communication Requirement that is integrated into both the HASS Requirement and the requirements of each major; see details below.

Summary of Subject Requirements	Subjects
Science Requirement	6
Humanities, Arts, and Social Sciences (HASS) Requirement [between one and three subjects can be from the Departmental Program]; at least two of these subjects must be designated as communication-intensive (CI-H) to fulfill the Communication Requirement.	8
Restricted Electives in Science and Technology (REST) Requirement [can be satisfied by 6.042(J) and 18.06 in the Departmental Program]	2

Figura: Uma nova graduação no MIT.

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

Visite:

www.analisemacro.com.br

