Perspectivas e oportunidades para estatísticos e estatísticas na emergente ciência de dados

Cibele Russo

ICMC USP

20/10/2022 - Semana da Estatística - UFAM

Sobre mim

- Bacharelado em Matemática Aplicada e Computação Científica (ICMC USP, 2005)
- Mestrado em Ciências de Computação e Matemática Computacional (ICMC USP, 2006)
- Doutorado em Estatística (IME USP, 2010)
- Professora de Estatística, ex-Coordenadora do Bacharelado em Estatística e Ciência de Dados e Professora do MBA em Ciências de Dados (ICMC USP)

Algumas experiências selecionadas

- 2004: Estatcamp - Consultoria Estatística em Qualidade, São Carlos SP
- 2006: Itaú - Itaú/Unibanco, São Paulo SP
- 2008: IPq - Hospital das Clínicas, São Paulo SP
- 2013: Biostatistics Dept, Erasmus Medical Center - Rotterdam, Netherlands
- 2022: Formação e Vida Profissional da Pró-Reitoria de Inclusão e Pertencimento da USP

Primeiro: Das dificuldades para definir o lugar da Ciência de Dados

- O que é um cientista de dados? Onde vive? O que come? :)
- "A data scientist is someone who is better at statistics than any software engineer and better at software engineering than any statistician." (Josh Wills ?)
- Frequentemente, o(a) estatístico(a) é descrito(a) como alguém que **domina os métodos e modelos teóricos** e o(a) cientista de dados é descrito(a) como alguém que **domina os algoritmos e métodos para a análise de dados**.

Como idealmente a Estatística faz parte da Ciência de Dados

- Ferramental teórico para a **formação sólida** do cientista de dados;
- **Compreensão e interpretabilidade** dos modelos que são populares hoje;
- **Criação de novos modelos**, de melhor performance que os existentes e mais adequados a cada problema;
- **Competitividade** do(a) cientista de dados.

Afinal de contas, o que é Ciência de Dados?

O que é Ciência? Ciência é Conhecimento. Ciência de dados seria conhecimento profundo sobre dados?

Como os dados se comportam? Como eles **variam**? Que **padrões** existem neles?

Data Science

"Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from noisy, structured and unstructured data, and apply knowledge and actionable insights from data across a broad range of application domains. Data science is related to data mining, machine learning and big data."
(https://en.wikipedia.org/wiki/Data_science)

Fonte: http://getc.com.tn/formation/big-data-data-science/

Data Science

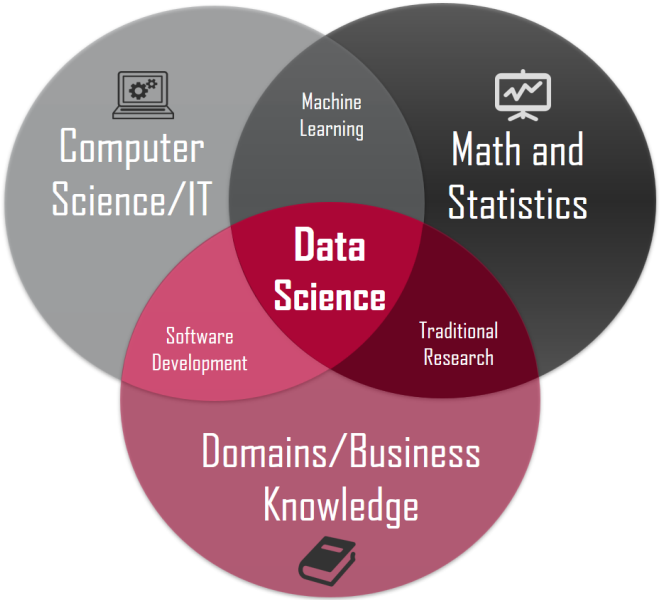
(https://en.wikipedia.org/wiki/Data_science)

Data science is a "concept to **unify statistics, data analysis, informatics, and their related methods**" in order to "understand and analyze actual phenomena" with data.

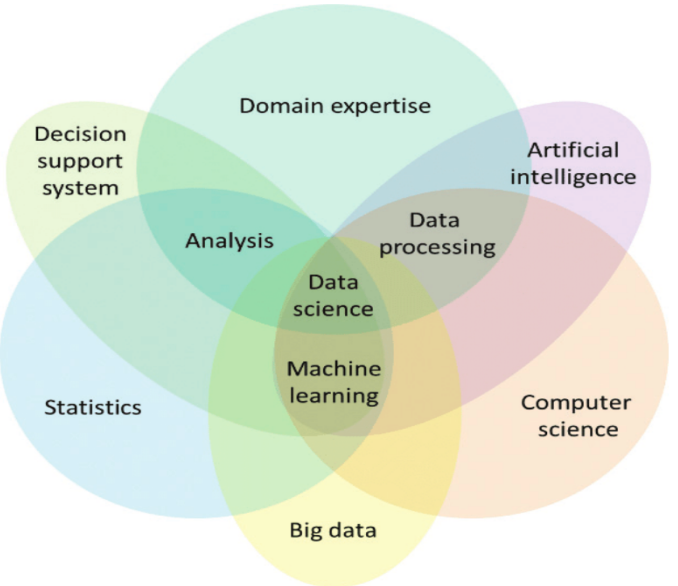
It uses techniques and theories drawn from many fields within the context of **mathematics, statistics, computer science, information science, and domain knowledge**.

However, data science is different from computer science and information science. Turing Award winner Jim Gray imagined **data science as a "fourth paradigm" of science (empirical, theoretical, computational, and now data-driven)** and asserted that **"everything about science is changing because of the impact of information technology"** and the data deluge.

Os famosos diagramas de Venn



Fonte: https://towardsdatascience.com/introduction-to-statistics-e9d72d818745



Fonte: https://www.researchgate.net/figure/The-field-of-data-science-including-statistics-big-data-and-artificial-intelligence-8_fig1_330948278

A Estatística e a Ciência de Dados

Estatística para Ciência de Dados, por Pedro A. Morettin e Júlio M. Singer

A Ciência de Dados e a Ciência dos Termos

APRENDIZADO (ESTATÍSTICO) DE MÁQUINA	ESTATÍSTICA
Redes, grafos (network, graphs)	Modelo (model)
Pesos (weights)	Parâmetros (parameters)
Atributos (Features/Inputs)	Covariáveis / Variáveis explicativas (covariates, explanatory variables)
Saída (Output)	Variável resposta / variável dependente (response variable, dependent variable)
Aprendizado (Learning)	Ajuste, Estimação (fitting, estimation)
Generalização (generalization)	Capacidade preditiva Avaliação em amostra teste (Performance test set)
Aprendizado Supervisionado (supervised learning)	Regressão (regression) Modelagem preditiva (predictive modeling)
Aprendizado Não Supervisionado (unsupervised learning)	Estimação de densidade, agrupamentos, redução de dimensionalidade. (density estimation, clustering, dimensionality reduction)

Machine learning: uma nova roupagem para a modelagem estatística

1. **Aprendizado supervisionado (Modelos de regressão)**
    - regressão: target ou valor alvo (variável resposta);
    - classificação: modelos em que a variável resposta é categórica;
2. **Aprendizado não-supervisionado (Modelos de análise multivariada)**
    - clustering (análise de agrupamentos);
    - análise de componentes principais;
    - análise de correspondência
3. **Aprendizado dinâmico (Modelos para séries de tempo)**
4. **Aprendizado profundo (pode ser visto com interpretação probabilística)**

... Outros

Aprendizado supervisionado (modelo de regressão)

Objetivos

Predizer Y a partir do conhecimento de variáveis em $X = x$.

Em notação matricial, um modelo linear geral é dado por

$$Y = X\beta + \epsilon,$$

em que

- Y é a **variável resposta** (vetor de variáveis aleatórias observáveis),
- X contém **variáveis preditoras** (matriz conhecida, ou seja, não-aleatória),
- β é um **vetor de parâmetros de interesse**, que queremos estimar,
- ϵ é o **erro aleatório** (vetor de erros aleatórios não observáveis).

Modelo de regressão linear

$$Y = X\beta + \epsilon,$$

em que

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, X = \begin{bmatrix} 1 & X_{11} & \dots & X_{1p} \\ 1 & X_{21} & \dots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \dots & X_{np} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

ou seja,

$$Y_{n\times 1} = X_{n\times(p+1)}\beta_{(p+1)\times 1} + \epsilon_{n\times 1}.$$

No modelo

$$Y = X\beta + \epsilon.$$

com as suposições

- $E(\epsilon) = 0$,
- $Var(\epsilon) = \sigma^2 I$,
- $(\epsilon \sim N(0, \sigma^2 I))$

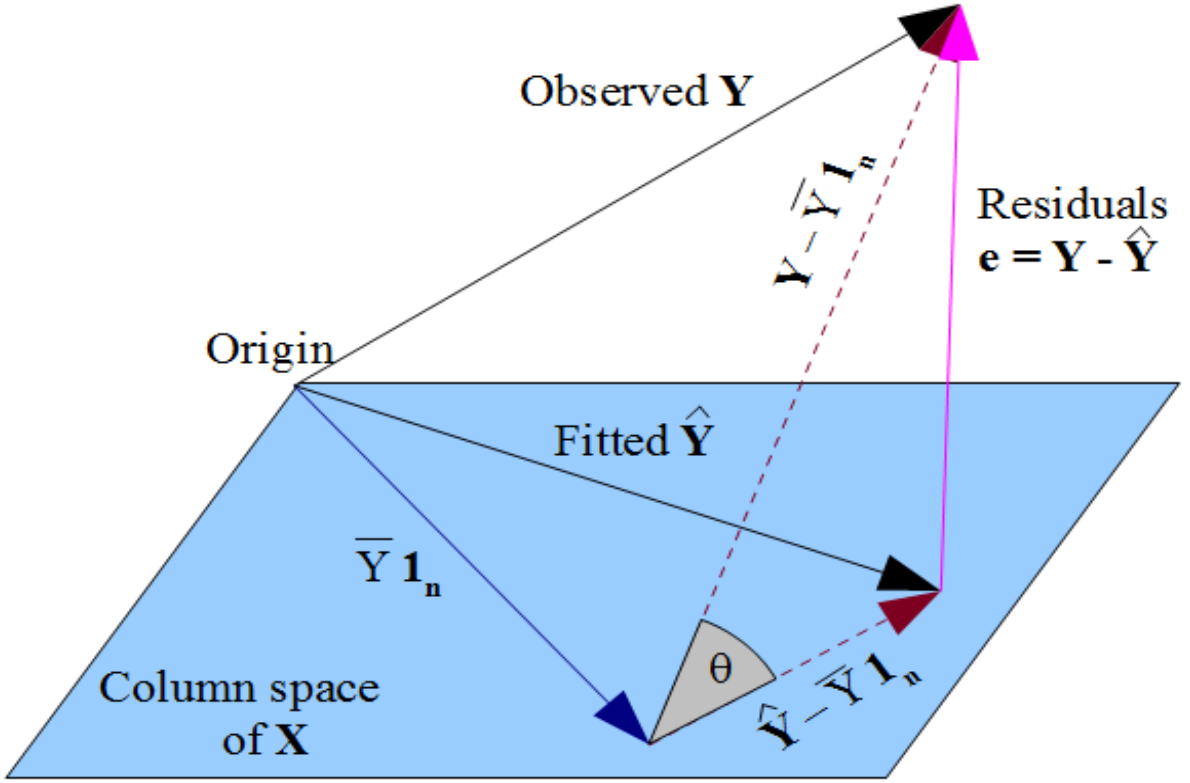
o estimador de mínimos quadrados (máxima verossimilhança) é dado por

$$\widehat{\beta} = (X^\top X)^{-1} X^\top Y.$$

O valor ajustado é

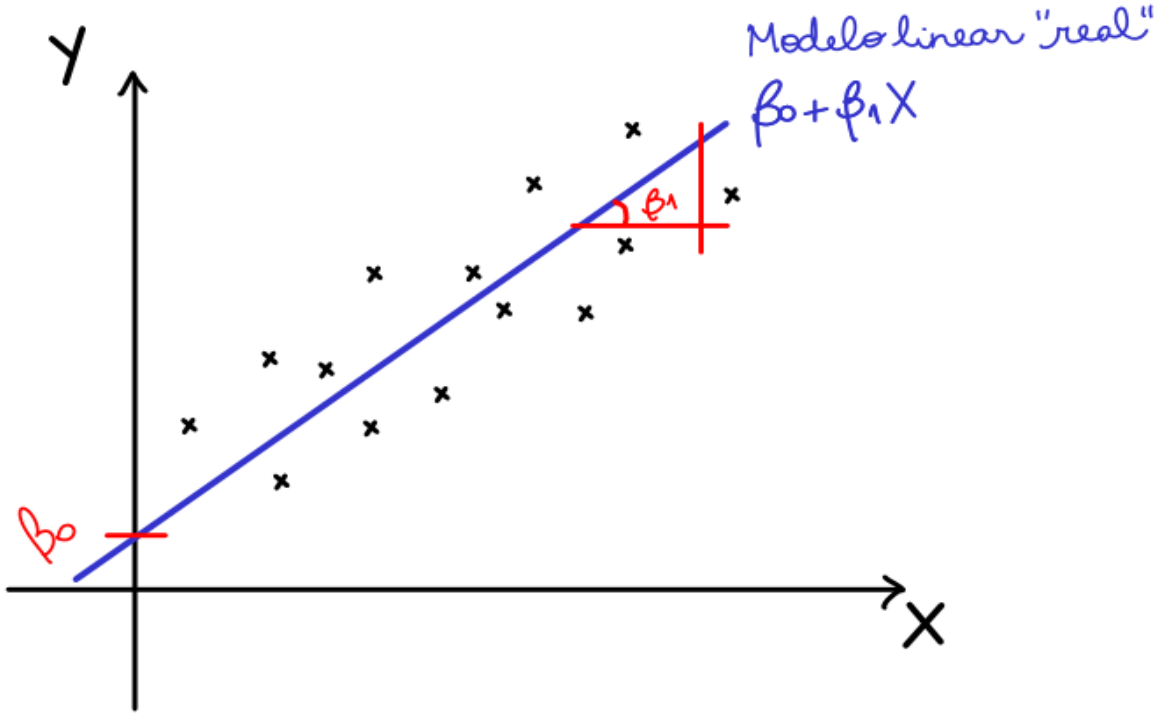
$$\widehat{Y} = X\widehat{\beta} = X(X^\top X)^{-1} X^\top Y = HY$$

Mas o que significa isso?

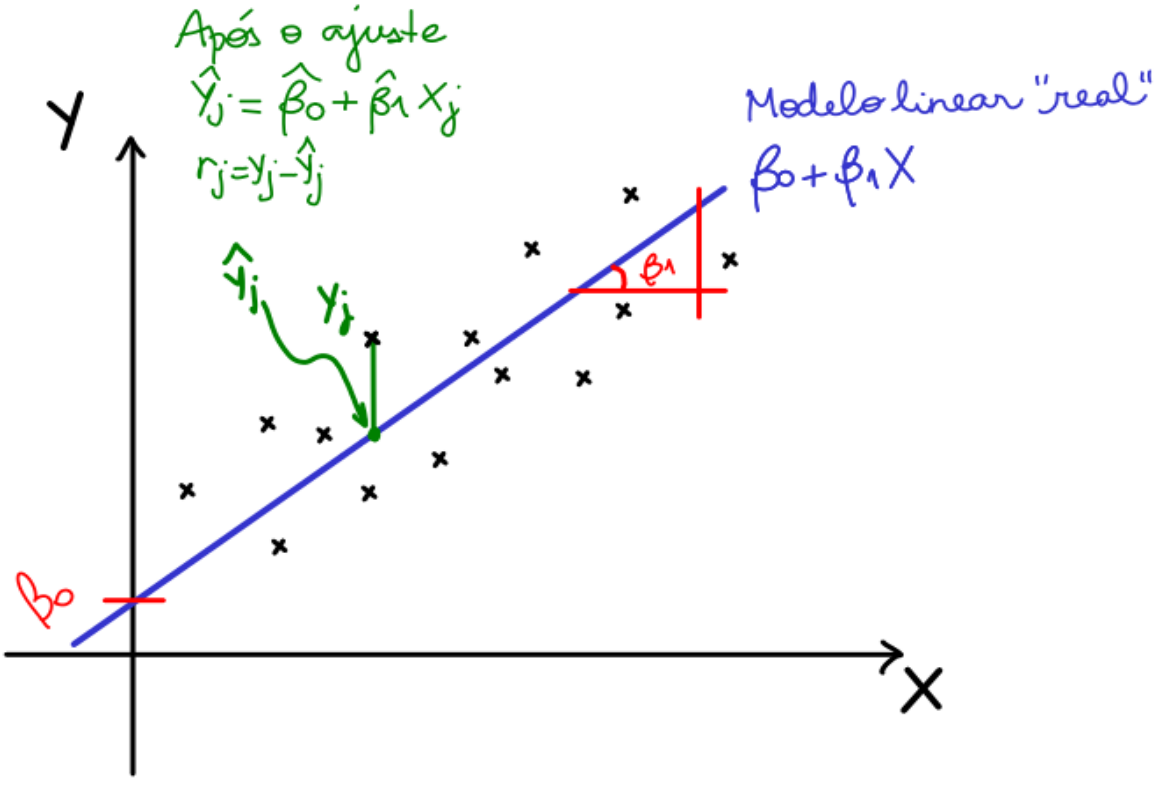


Fonte: https://stats.stackexchange.com/questions/123651/geometric-interpretation-of-multiple-correlation-coefficient-r-and-coefficient

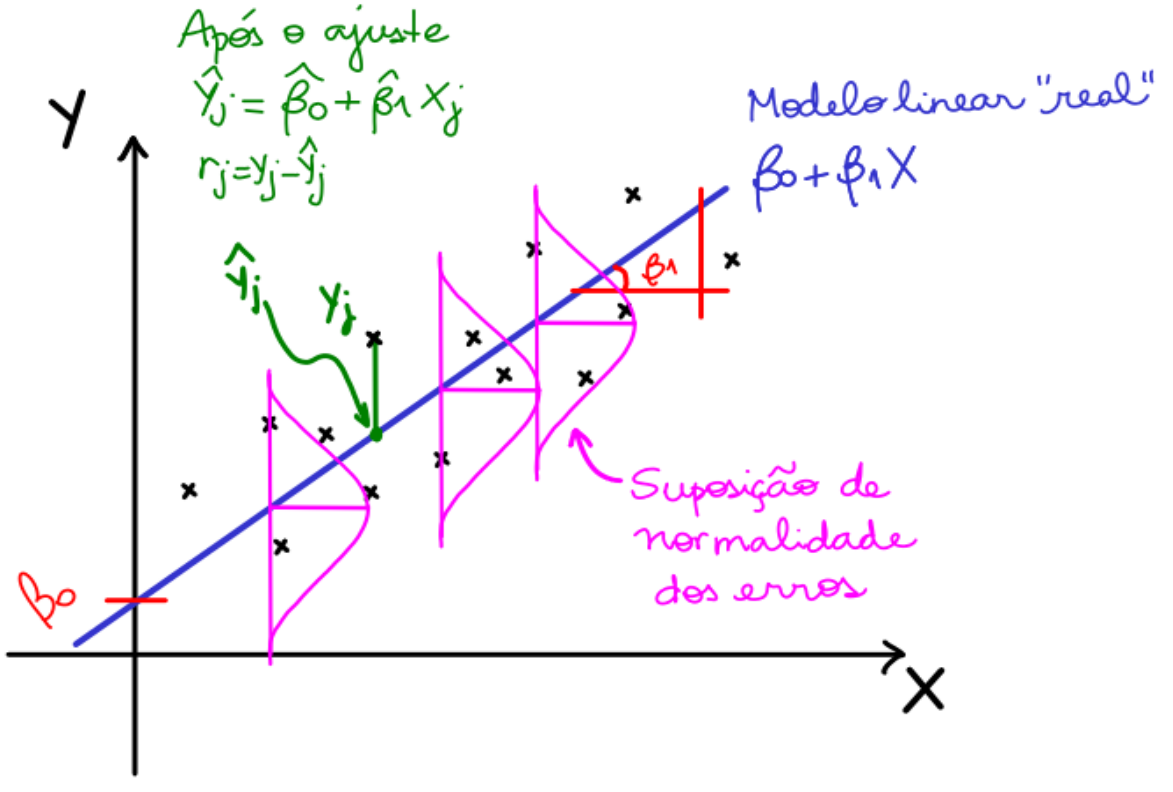
Vamos pensar no modelo de regressão linear simples.



Vamos pensar no modelo de regressão linear simples.



Vamos pensar no modelo de regressão linear simples.



E Aprendizado não-supervisionado?

Objetivo principal: Extrair informação útil de dados que não foram rotulados previamente.

Em outras palavras: Analisar dados sem uma variável específica que os rotule, seja ela quantitativa ou qualitativa.

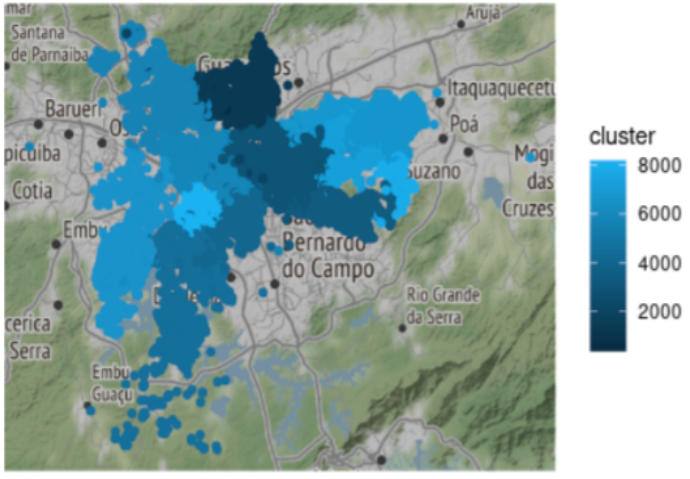


Figure 3. Nearest neighbors by default, with 33 clusters obtained

Exemplo agrupamentos. Fonte: https://www.cambridge.org/engage/miir/article-details/614501d66fc3a8bd14a2b6b2

Análise de Componentes Principais

A Análise de componentes principais (ACP ou PCA, de principal component analysis) é uma técnica que **transforma linearmente um conjunto de p variáveis correlacionadas em um conjunto de k variáveis não correlacionadas (com $k < p$)**, que explicam uma parcela substancial das informações do conjunto original.

Objetivos principais

- Reduzir a **dimensionalidade** dos dados.
- Obter **combinações interpretáveis** das variáveis originais.
- Descrever e compreender a **estrutura de correlação** das variáveis originais.

Contexto

Seja X um vetor aleatório de dimensão $p \times 1$ com vetor de médias (populacionais) $\mu_{p\times 1}$ e matriz de variâncias e covariâncias (populacionais) de $\Sigma_{p\times p}$.

Estamos particularmente interessados no caso em que as variáveis $X_1, \ldots, X_p$ estão correlacionadas, isto é, algumas (ou muitas) das covariâncias $Cov(X_i, X_j), i, j = 1, \ldots, p$ e $i \neq j$ são não-nulas.

Quando isso ocorre, significa que existe **redundância entre dimensões**, e podemos ter interesse em reduzir a dimensionalidade do problema, construindo novas variáveis, não correlacionadas entre si, que sejam combinações lineares das variáveis originais.

Pode ser que **poucas (k) novas variáveis expliquem grande parte da variabilidade existente nas (p) variáveis originais**. Isso pode significar a **redução de custos como tempo computacional e espaço para armazenamento de dados**.

Análise de componentes principais

Seja $X \sim (\mu, \Sigma)$. Sejam $\lambda_1 \geq \ldots \geq \lambda_p$ os autovalores de Σ, com autovetores correspondentes $e_1, \ldots, e_p$, tais que

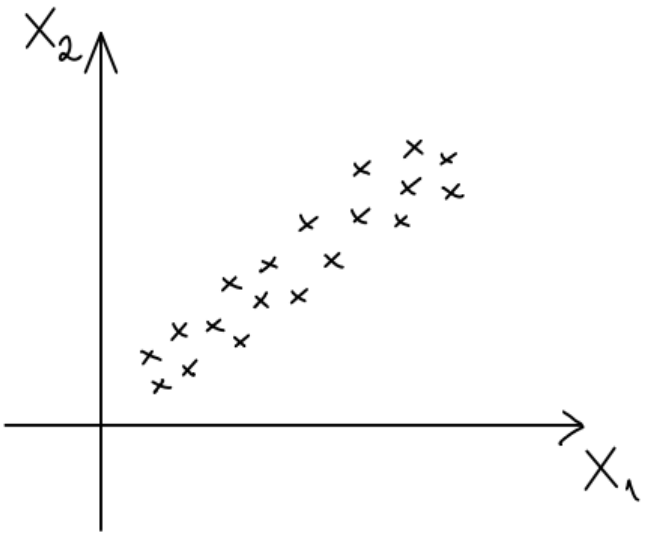
- $e_i^\top e_j = 0$, para $i, j = 1, \ldots, p$ e $i \neq j$,
- $e_i^\top e_i = 1$, para $i = 1, \ldots, p$,
- $\Sigma e_i = \lambda_i e_i$, para $i = 1, \ldots, p$.

Considere a matriz ortogonal $O_{p \times p} = (e_1, \ldots, e_p)$.

Então $Y_{p \times 1} = O^\top X$ é o vetor de componentes principais de Σ.

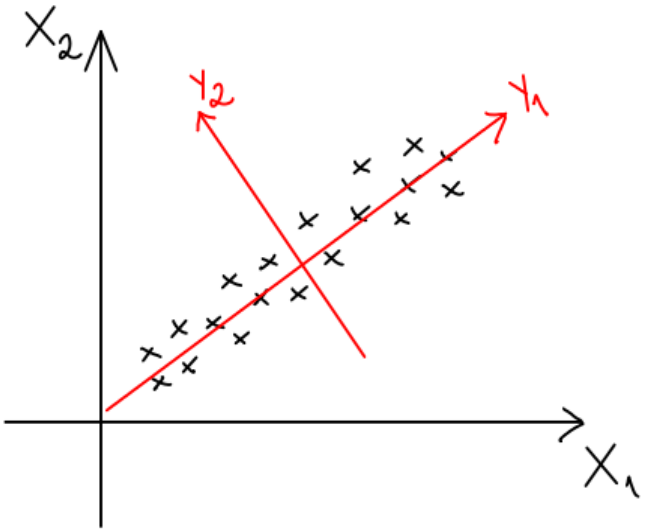
Análise de componentes principais

Interpretação geométrica (p=2)



Análise de componentes principais

Interpretação geométrica (p=2)



Análise de componentes principais

Propriedades:

- A j-ésima componente principal de Σ é dada por

$$Y_j = {e_j}^\top X.$$

- $E(Y_j) = {e_j}^\top \mu$.
- $Var(Y_j) = {e_j}^\top \Sigma e_j = \lambda_j$.
- $Cov(Y_i, Y_j) = Cor(Y_i, Y_j) = 0$ para $i, j = 1, \ldots, p$ e $i \neq j$.
- A proporção da variância total de X que é explicada pela j-ésima componente principal é

$$\frac{\lambda_j}{\sum_{i=1}^{p} \lambda_i}.$$

Estimação das componentes principais

Como em geral a matriz Σ é desconhecida, utiliza-se a matriz S, de variâncias e covariâncias amostrais, para **estimar** as componentes principais.

Considere $\hat{\lambda}_1 \geq \ldots \geq \hat{\lambda}_p$ os autovalores de S, com autovetores correspondentes padronizados $\hat{e}_1, \ldots, \hat{e}_p$.

A j-ésima componente principal amostral é dada por

$$\hat{Y}_j = \hat{e}_j^\top X.$$

O que isso significa?

Propriedades das componentes principais

- $\widehat{Var}(\hat{Y}_j) = \hat{\lambda}_j$.
- $Cov(\hat{Y}_i, \hat{Y}_j) = Cor(\hat{Y}_i, \hat{Y}_j) = 0$ para $i, j = 1, \ldots, p$ e $i \neq j$.
- A proporção da variância total explicada pela j-ésima componente principal amostral é

$$\frac{\hat{\lambda}_j}{\sum_{i=1}^{p} \hat{\lambda}_i}$$

- $\widehat{Cor}(\hat{Y}_j, X_i) = \dfrac{\hat{e}_{ji}\sqrt{\hat{\lambda}_j}}{\sqrt{s_{jj}}}$

Estimação das componentes principais

- Pelo teorema da decomposição espectral,

$$S_{p\times p} = \sum_{j=1}^{p} \hat{\lambda}_j \hat{e}_j \hat{e}_j^\top$$

  pode ser aproximada por

$$S_{p\times p} \approx \sum_{j=1}^{k} \hat{\lambda}_j \hat{e}_j \hat{e}_j^\top$$

Escores das componentes

É comum utilizar os **escores das componentes** para análises estatísticas ou ordenação (**ranking**) dos elementos amostrais.

Esses escores são obtidos ordenando os valores obtidos de

$$\hat{Y}_j = \hat{e}_j^\top X.$$

Mais em:
https://github.com/cibelerusso/AnaliseMultivariadaEAprendizadoNaoSupervisiona

CD: uma ameaça para a estatística?

(Data Scientists Versus Statisticians https://medium.com/odscjournal/data-scientists-versus-statisticians-8ea146b7a47f em tradução livre)

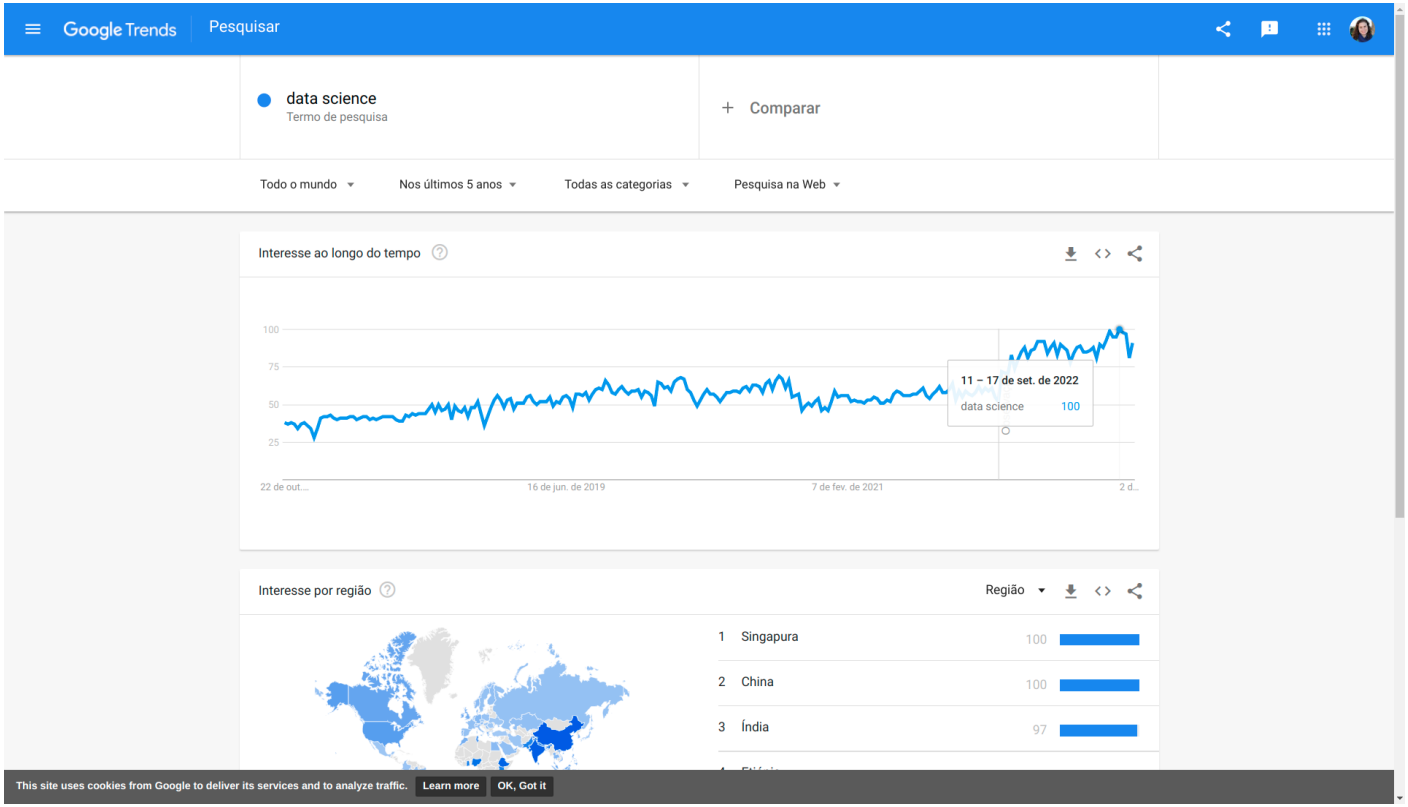
"Não é de se admirar que os estatísticos se sintam ameaçados pelos cientistas de dados em algum nível. Os estatísticos lidam com conceitos **nebulosos** como estimativas pontuais, margens de erro, intervalos de confiança, erros padrão, valores-p, teste de hipóteses e o argumento proverbial entre os “frequentistas” e os “bayesianos”.

Os estatísticos podem ser vistos como confusos para o público em geral e muitas vezes os estatísticos nem conseguem concordar com o que é correto."

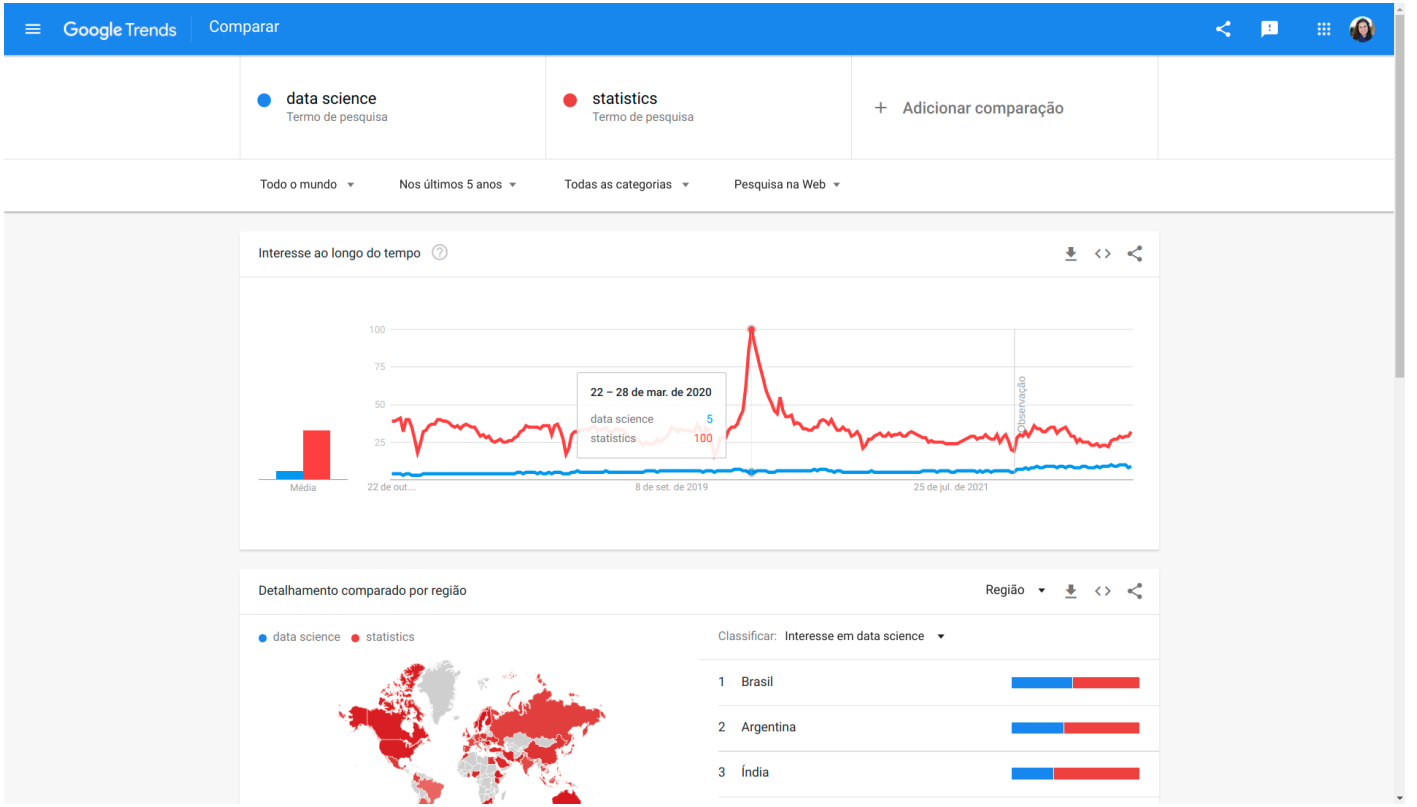
"Os cientistas de dados, por outro lado, seguem de perto o “processo de ciência de dados” que é mais acessível; ingestão de dados, transformação de dados, análise exploratória de dados, seleção de modelos, avaliação de modelos e narrativa de dados. Claro, muitas dessas etapas seguem métodos estatísticos nos bastidores, mas são seladas em um invólucro mais envolvente e compreensível.

Muito mais pessoas podem adotar a ciência de dados."

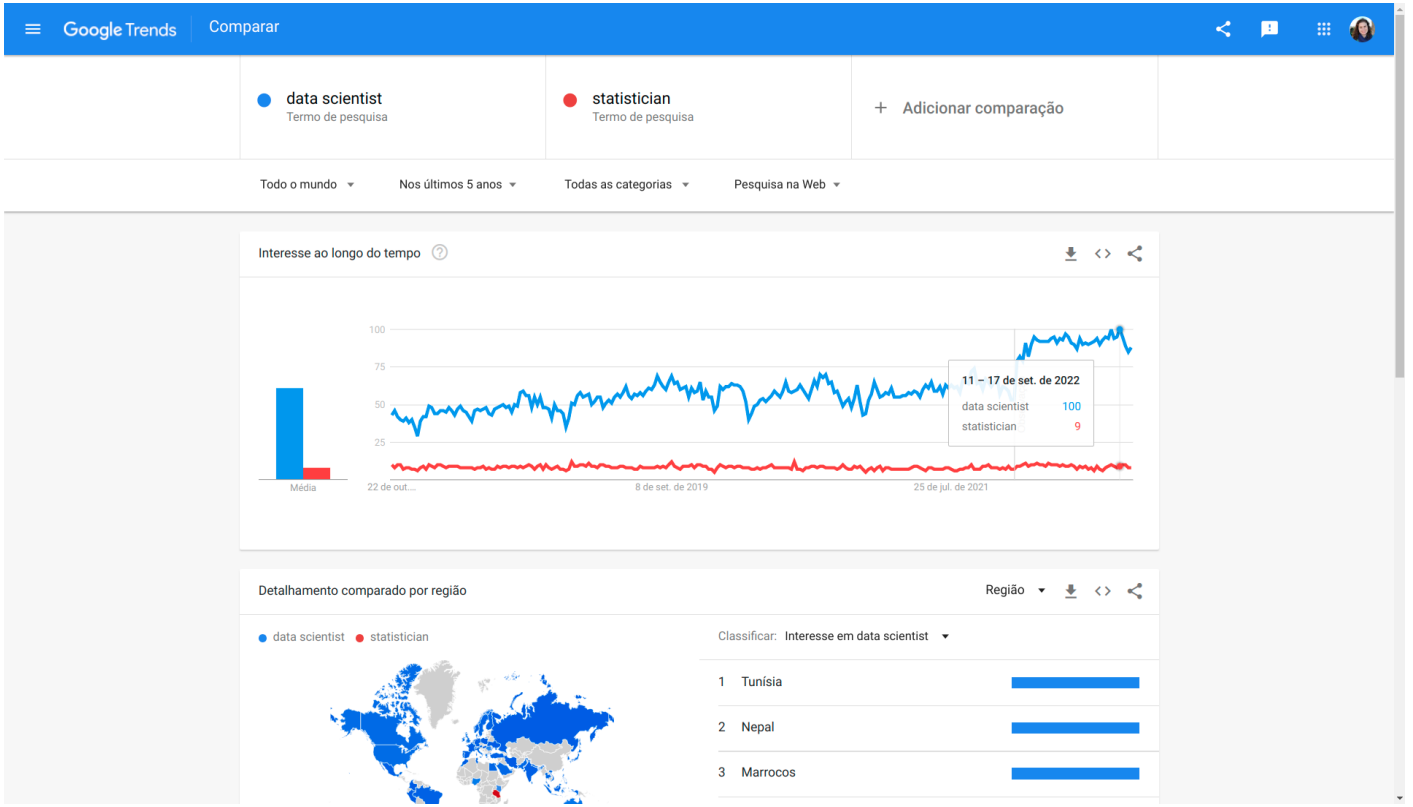
Buscas por 'data science'



Buscas por 'data science' e 'statistics'

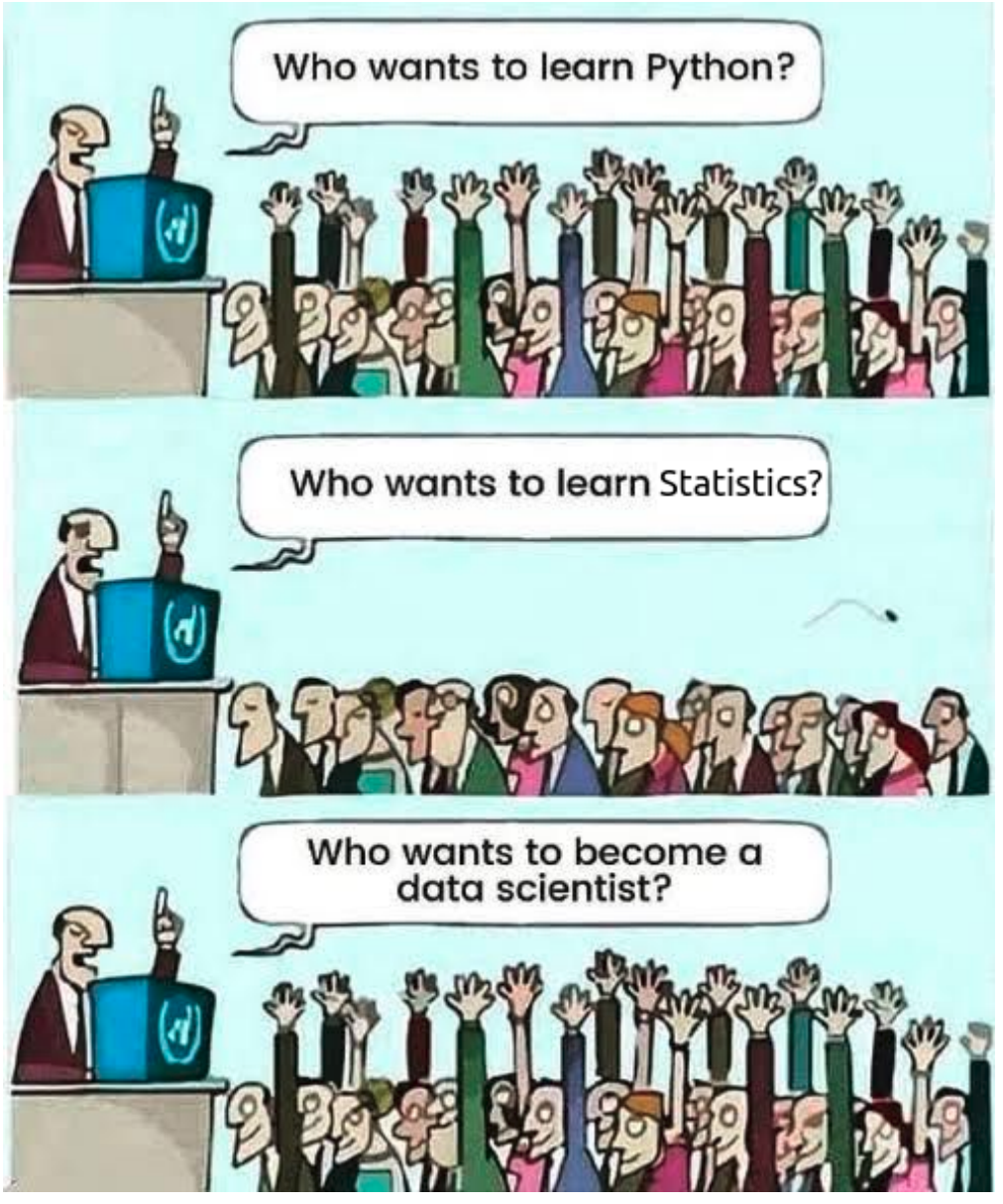


Buscas por 'data scientist' e 'statistician'



E onde se coloca o profissional da Estatística dentro do mercado de Ciência de Dados?

- Domínio da teoria estatística para a ciência de dados;
- Ferramentas inferenciais e probabilísticas;
- Domínio de técnicas de amostragem;
- Modelagem de dados e diagnóstico de modelos;
- Interpretabilidade de resultados;
- ....



Fonte: incerta

Oportunidades e desafios para estatísticos e estatísticas na ciência de dados

- O que mais preciso estudar?
- Como posso aprimorar minhas **habilidades computacionais**?
- R OU Python? (ou outras linguagens? Por que não? Por que não R E Python?)
- Como aprimorar a **comunicação, disponibilização e mesmo propaganda** dos métodos estatísticos? (github? Medium?)

Para onde seguir depois?

- Especialização? MBA?
- Mestrado?

Excelentes opções oferecidas pelo ICMC

Uma introdução ao Python para Estatística e Ciência de Dados:

https://bit.ly/3Tny9ce