



Análise de Dados com Python



Programa do curso

Segunda: Conceitos iniciais (Manhã)

Python para manipular tabelas (Tarde)

Terça: Python para manipular tabelas

Quarta: Visualizando os dados em gráficos

Quinta: Regressão Linear (Manhã)

Árvores de Decisão (Tarde)

Sexta: Aplicações avançadas (Manhã)

Plataforma de estudos do ODP (Tarde)



Instrutores

**Cláudio
Fernando**



**Gustavo
Ricardo
Rodrigo**



Instrutores

**Cláudio
Fernando**



**Gustavo
Ricardo
Rodrigo**





Análise de Dados com Python

Conceitos

Ricardo Silva Carvalho
ODP/DIE

Dados: o novo petróleo

[1] São combustível para a ciência de dados

Ciência de dados converte dados em algo útil para tomada de decisão



[2] Aproveitamento vem após longo tratamento (refino)

Necessário para o real uso



[3] Já existia antes, mas não tão aproveitado

Após investimento em obtenção



Dados? Quero!

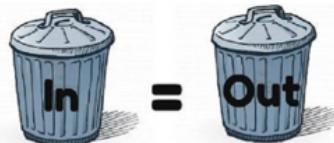
“Em Deus nós confiamos. Os outros tragam dados.”



Caos!

Dados crescendo exponencialmente: Sistema caótico

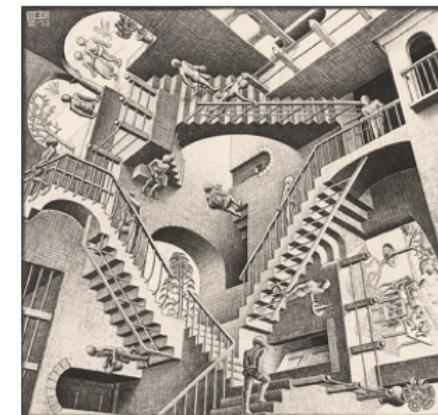
Garbage in → Garbage out



À medida que dados tornam-se mais comuns
não é quantidade que definirá o **valor** dos dados,
mas **sim a qualidade!**

"Adoramos o caos porque adoramos produzir ordem."

– M. C. Escher



QUALIDADE > QUANTIDADE

É um ou outro?

Mas meus dados são não estruturados e mal organizados!!!

"É por esse motivo que apenas 0,5% dos dados empresariais são analisados e usados." – Forbes

Seria um problema de visão?

Dados: Futuro?

Cenário já mais comum atualmente:

Análise de dados somente como útil para gerar valor de dados existentes

Análise de dados como **Ativo Estratégico**:

Possuir os dados certos
Construir talentos

→ São complementares!

Não é trivial e necessita de investimentos!

Dados como Ativo estratégico

Exemplo na área bancária:

[1980] Risco de crédito: Probabilidade de inadimplência
Cartões de crédito tinham preços uniformes



[1990] Fairbanks e Morris: medir lucratividade e dar condições diferenciadas
Ideia era ter diferentes preços, limites, taxas, etc.

[PROBLEMA]: Ninguém tinha os dados necessários para modelagem
Só tinham dados das condições fixas e para baixa prob. de inadimplência.
Somente um pequeno banco da Virginia aceitou: **Signet Bank**

[RESULTADO]: Saíram de 2,9% de inadimplência para 6%

Dados como Ativo estratégico

Anos se passaram.. Muitos testes.. Investimentos continuaram..
Dados certos foram coletados e analisados..

Deu certo!

→ Fairbanks e Morris se tornaram CEO e COO

Signet-Bank?

Empresa spin-off: **Capital One**

Hoje uma das maiores emissores de cartão de crédito com menor taxa de inadimplência

Em 2000 divulgaram que estavam em torno de 45000 “testes científicos”



Quais dados?

“You can’t manage what you don’t measure.”

– W. Edwards Deming

1ª Pergunta: Você conhece seus dados?

2ª Pergunta: Dados que tem são fotografia ou
são atualizados frequentemente?

Tipos de Dados

Identificar problemas solucionáveis usando dados

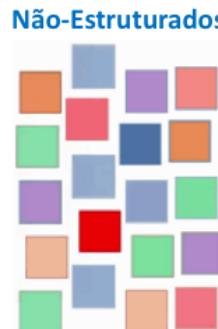
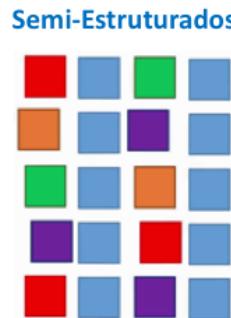
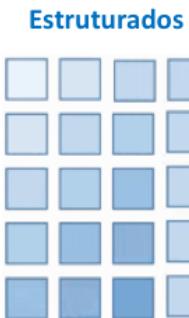
Tipos (básicos) de Dados

- Número Inteiro: 1, 2, 3, 4
- Número Real/Decimal/Ponto Flutuante: 1.70, 2.10, 9.99
- Binário/Booleano/Lógico: 0 ou 1, TRUE ou FALSE
- Caractere/String/Texto/Categórico: "Fulano", "Brasília", 'CGU', 'DAS-99'



Tipos de Dados

Estruturados vs Semi-Estruturados vs Não Estruturados



Tipos (básicos) de Dados



DAS	Idade
3	32
2	29
2	31
4	39
1	27
5	44
2	30
3	33

Tipos dos Dados

Idade: Numérico

DAS: Numérico

Algum problema nisso?

DAS-3 é três vezes mais importante que DAS-1?

Dado numérico nem sempre pode ser tratado/mantido como numérico

Tipos de Dados

Estruturados vs Semi-Estruturados vs Não Estruturados

Age	Has_Job	Own_House	Credit_Rating	Class
young	false	false	fair	No
young	false	false	good	No
young	true	false	good	Yes
young	true	true	fair	Yes
young	false	false	fair	No
middle	false	false	fair	No
middle	false	false	good	No
middle	true	true	good	Yes
middle	false	true	excellent	Yes
middle	false	true	excellent	Yes
old	false	true	excellent	Yes
old	false	true	good	Yes
old	true	false	good	Yes
old	true	false	excellent	Yes
old	false	false	fair	No

Estruturados

Tipos de Dados

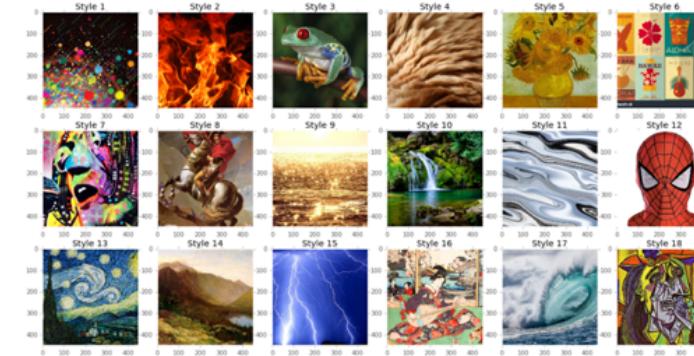
Estruturados vs Semi-Estruturados vs Não Estruturados

```
{  
    "custkey": "450002",  
    "useragent": {  
        "devicetype": "pc",  
        "experience": "browser",  
        "platform": "windows"  
    },  
    "pagetype": "home",  
    "productline": "television",  
    "customerprofile": {  
        "age": 20,  
        "gender": "male",  
        "customerinterests": [  
            "movies",  
            "fashion",  
            "music"  
        ]  
    }  
}
```

Semi-Estruturados

Tipos de Dados

Estruturados vs Semi-Estruturados vs Não Estruturados



Não Estruturados

Tipos de Dados

Estruturados vs Semi-Estruturados vs Não Estruturados

```
<?xml version="1.0" encoding="UTF-8"?>  
<breakfast_menu>  
    <food>  
        <name>Belgian Waffles</name>  
        <price>$5.95</price>  
        <description>Two of our famous Belgian Waffles with plenty of real maple syrup.</description>  
        <calories>650</calories>  
    </food>  
    <food>  
        <name>Strawberry Belgian Waffles</name>  
        <price>$7.95</price>  
        <description>Light Belgian waffles covered with strawberries and whipped cream.</description>  
        <calories>650</calories>  
    </food>  
    <food>  
        <name>Berry-Berry Belgian Waffles</name>  
        <price>$8.95</price>  
        <description>Light Belgian waffles covered with an assortment of fresh berries and  
        whipped cream.</description>  
        <calories>900</calories>  
    </food>  
</breakfast_menu>
```

Semi-Estruturados

Tipos de Dados

Rotulados vs Não Rotulados

Age	Has_Job	Own_House	Credit_Rating	Class
young	false	false	fair	No
young	false	false	good	No
young	true	false	good	Yes
young	true	true	fair	Yes
young	false	false	fair	No
middle	false	false	fair	No
middle	false	false	good	No
middle	true	true	good	Yes
middle	false	true	excellent	Yes
middle	false	true	excellent	Yes
old	false	true	excellent	Yes
old	false	true	good	Yes
old	true	false	good	Yes
old	true	false	excellent	Yes
old	false	false	fair	No

Estruturados Rotulados

Tipos de Dados

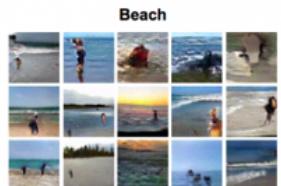
Rotulados vs Não Rotulados

worst area	worst smoothness	worst compactness	worst concavity	worst concave points	worst symmetry	worst fractal dimension	target
677.9	0.14260	0.2378	0.2671	0.10150	0.3014	0.08750	benign
1866.0	0.11930	0.2336	0.2687	0.17890	0.2551	0.06589	malignant
1156.0	0.15460	0.2394	0.3791	0.15140	0.2837	0.08019	malignant
515.8	0.14500	0.2629	0.2403	0.07370	0.2556	0.09359	benign
457.8	0.13450	0.2118	0.1797	0.06918	0.2329	0.08134	benign

Estruturados e Rotulados

Tipos de Dados

Rotulados vs Não Rotulados



Não Estruturados
Rotulados



Tipos de Dados

Rotulados vs Não Rotulados



Não Estruturados
Não Rotulados

Tipos de Dados

Estruturados vs Semi-estruturados vs Não estruturados

Rotulado vs Não rotulado

- Rótulo (Label): Usado para predição ou reconhecimento (Categórico ou Numérico)
- Aprendizagem **Supervisionada** vs Aprendizagem **Não Supervisionada**

Dados = Atributos/Features/Var. Indep. + Rótulo/Label/Alvo/Classe (opcional)

worst area	worst smoothness	worst compactness	worst concavity	worst concave points	worst symmetry	worst fractal dimension	target
677.9	0.14260	0.2378	0.2671	0.10150	0.3014	0.08750	benign
1866.0	0.11930	0.2336	0.2687	0.17890	0.2551	0.06589	malignant
1156.0	0.15460	0.2394	0.3791	0.15140	0.2837	0.08019	malignant
515.8	0.14500	0.2629	0.2403	0.07370	0.2556	0.09359	benign
457.8	0.13450	0.2118	0.1797	0.06918	0.2329	0.08134	benign

→ Label/Rótulo/Alvo

Exemplo: Dados e Classificação



Categórico (Chave)	Categórico (Constante)	Categórico	Real	Categórico	Binário	Classe
330-272-449	Seaborg	Good	0.123	red	1	Yes
330-272-450	Seaborg	Bad	0.987	green	1	No
330-272-451	Seaborg	Yes	0.245	blue	0	Yes
720-273-500	Seaborg	Yes	0.254	blue	1	Yes
720-273-501	Seaborg	Bad	0.244	blue	0	No
720-273-502	Seaborg		0.415	green	0	Maybe
110-272-461	Seaborg	Yes	0.925	red	1	Yes
110-272-462	Seaborg	Yes	0.376	green	0	Yes
220-273-700	Seaborg	Bad	0.615	green	1	No
220-274-701	Seaborg		0.321	blue	0	Maybe
220-275-703	Seaborg	Bad	0.098	green	0	No
220-275-704	Seaborg	Bad	0.765	red	1	No

PROXY da Classe

Exemplo: Dados e Classificação



Exemplo: Dados e Classificação



Dados = Atributos + Classe

Real	Categórico	Binário	Classe
0.123	red	1	Yes
0.987	green	1	No
0.245	blue	0	Yes
0.254	blue	1	Yes
0.244	blue	0	No
0.415	green	0	Maybe
0.925	red	1	Yes
0.376	green	0	Yes
0.615	green	1	No
0.321	blue	0	Maybe
0.098	green	0	No
0.765	red	1	No

Ninguém vai limpar para mim?



Data Warehouses vs Data Lake vs Data Swamp

“Um data lake é um repositório que contém uma grande quantidade de dados brutos em seu formato nativo, incluindo dados estruturados, semiestruturados e não estruturados. A estrutura de dados e os requisitos não são definidos até que os dados sejam necessários/usados”

Dados: Modelados **vs** Qualquer (Schema-on-write **vs** Schema-on-read)

Agilidade: Mudança de estrutura **vs** Nenhuma estrutura

Usuários: Venham todos **vs** Cientistas de dados

Ambientes



Data Warehouses vs Data Lake vs Data Swamp



Tipos de Análise de dados

Tipos de Análise de Dados

Em ordem aproximada de dificuldade:

- Descritiva
- Exploratória
- Inferencial
- Preditiva
- Causal
- Mecanicista



Análise Descritiva

OBJETIVO: Descrever um conjunto de dados

Média $\bar{x} = \frac{\sum x_i}{N}$

Moda Mais frequente

Mediana Divide dados em dois



EXEMPLO:

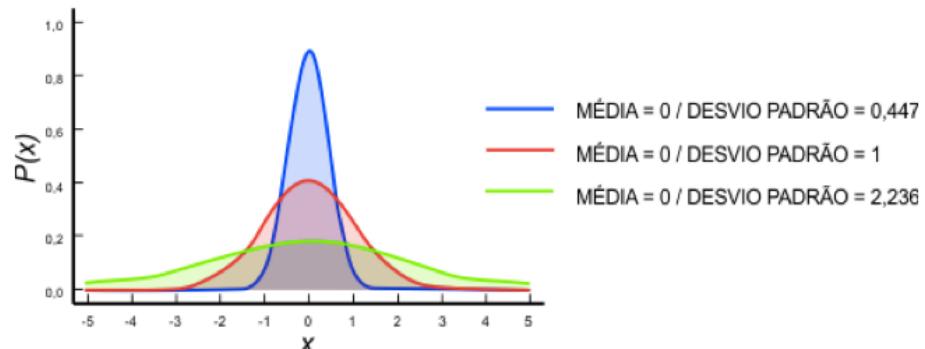
Dias sem pagar cartão → 8, 2, 5, 10, 4, 12, 2, 2, 81

Média → $126/9 = 14$ Moda → 2

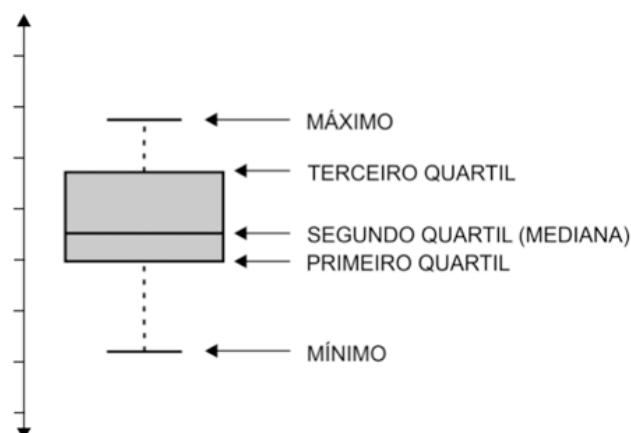
Mediana → 2, 2, 2, 4, 5, 8, 10, 12, 81

Análise Descritiva

$$\text{Variância (amostral)} \quad s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

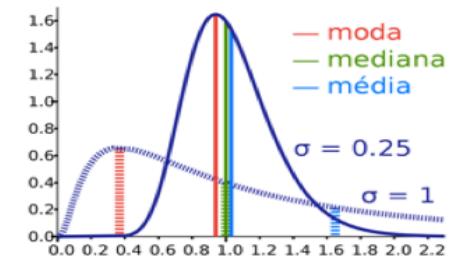


Análise Descritiva



Análise Descritiva

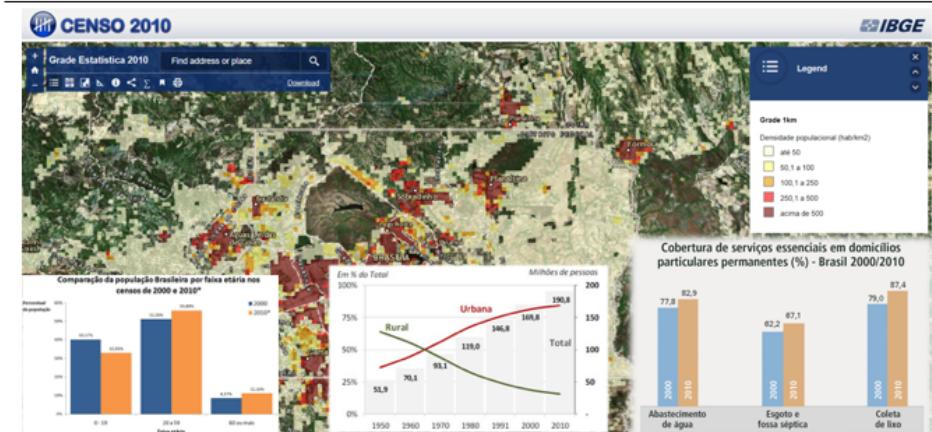
- Idade de dependente:
2, 4, 5, 7, 8, 10, 12, 13, 83
- Média = $144/9 = 16$
83 influencia distorcendo a média
- Mediana = 8



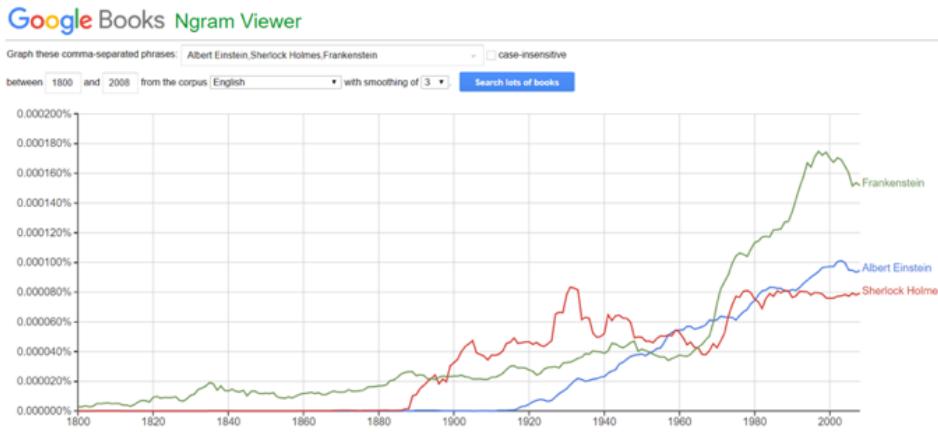
MAIS UM EXEMPLO PRÁTICO:

Cidade com 1000 habitantes com média e mediana de \$ 80 mil anuais
Warren Buffet decide ir morar na cidade → \$ 1 bilhão anualmente
Mediana continua \$ 80 mil, mas média fica em \$ 1,08 milhão
Cidade vira “A cidade dos milionários” por causa de um habitante?

Análise Descritiva



Análise Descritiva



Análise Descritiva

Automobile Price Dataset

	make	fuel.type	aspiration	num.of.doors
toyota	: 32	diesel: 20	std :168	Min. :2.000
nissan	: 18	gas :185	turbo: 37	1st Qu.:2.000
mazda	: 17			Median :4.000
honda	: 13			Mean :3.123
mitsubishi	: 13			3rd Qu.:4.000
subaru	: 12			Max. :4.000
(Other)	: 100			NA's :2

	VALOR	TAXA
count	29903.00	29903.00
mean	602.09	2.03
std	365.51	38.55
min	0.00	0.00
25%	333.84	0.00
50%	517.71	0.00
75%	740.02	0.00
max	3575.03	1760.90

Passagens Aéreas

Análise Exploratória

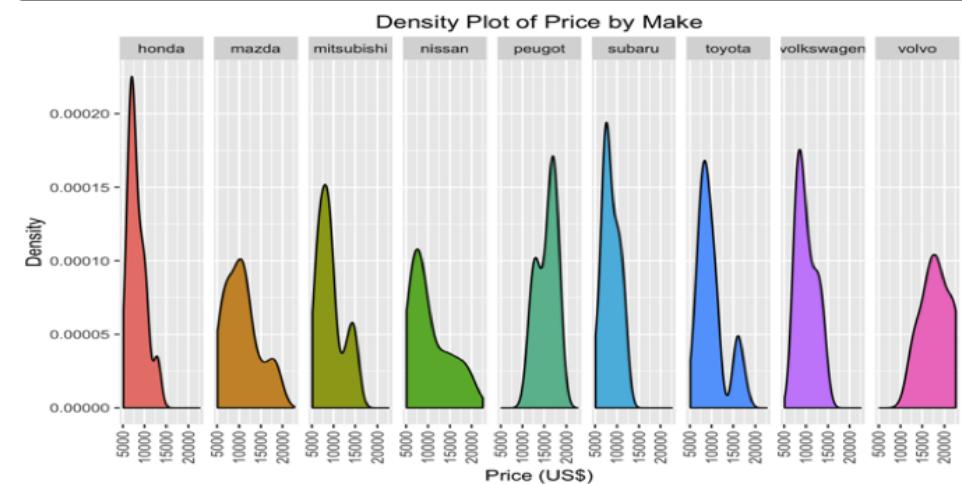
OBJETIVO: Encontrar relações/padrões antes desconhecidos

- Analisar entendimento prévio/preconcebido
- Abrir novos caminhos ou estratégias



Isoladamente **não** deve ser usada para generalizar/prever

Análise Exploratória



Análise Exploratória

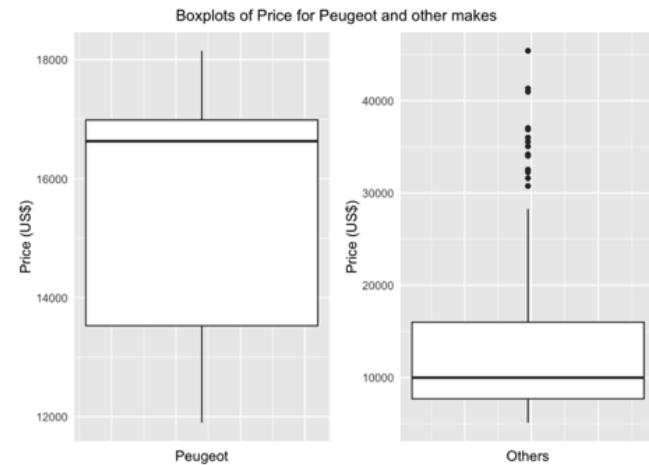


Gráfico não é tão importante...

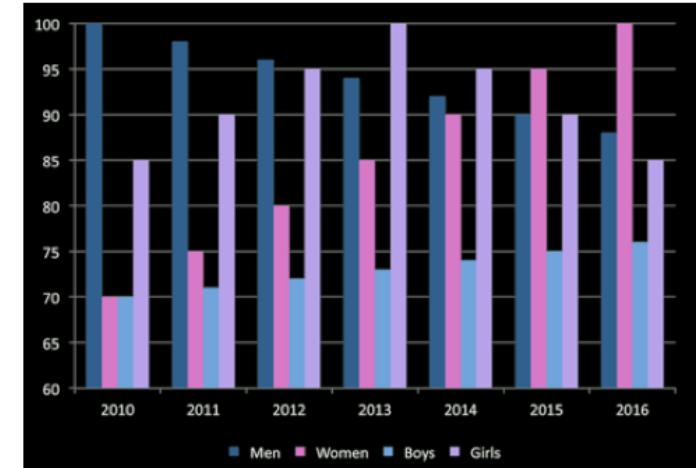
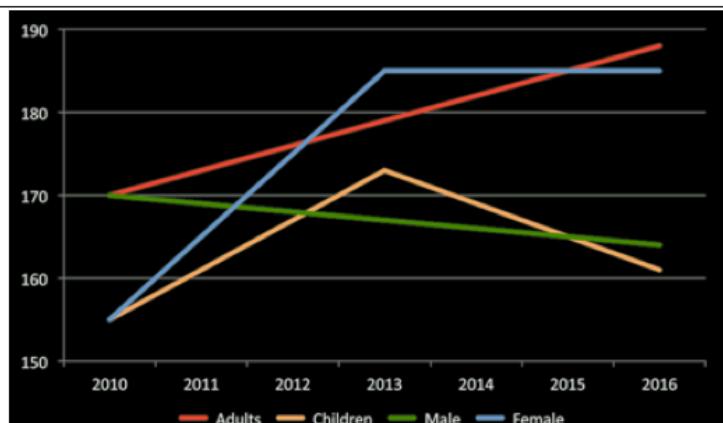
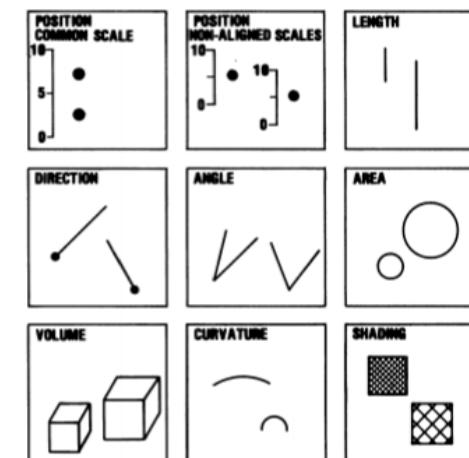


Gráfico não é tão importante...



"Dados torturados confessam"

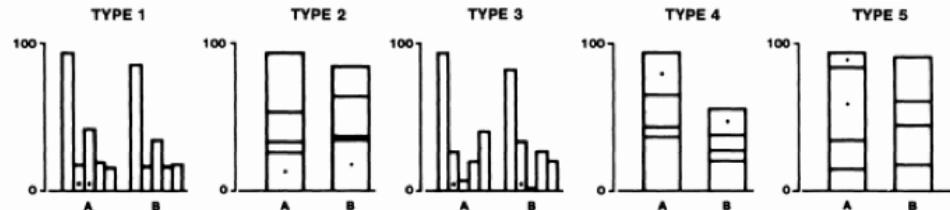
Gráfico é tranquilo...



[Graphical perception: Theory, Experimentation, and Applications to the Development of Graphical Models](#)

Gráfico é tranquilo...

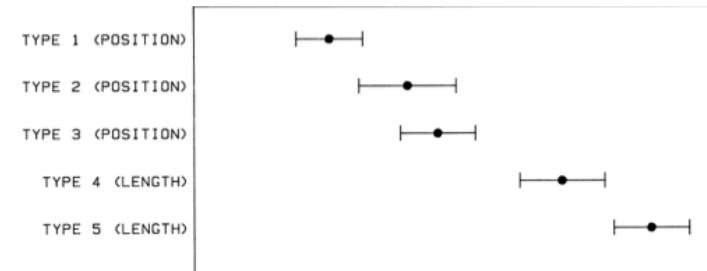
Posição vs Tamanho



[Graphical perception:
Theory, Experimentation,
and Applications to the
Development of Graphical
Models](#)

Gráfico é tranquilo...

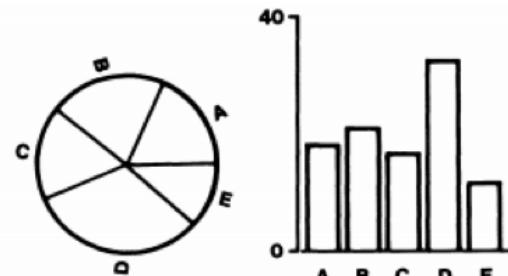
Posição vs Tamanho



[Graphical perception:
Theory, Experimentation,
and Applications to the
Development of Graphical
Models](#)

Gráfico é tranquilo...

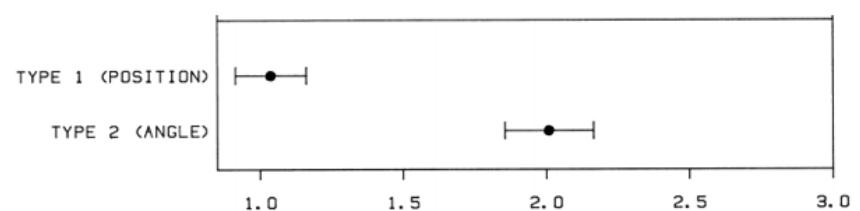
Posição vs Ângulo



[Graphical perception:
Theory, Experimentation,
and Applications to the
Development of Graphical
Models](#)

Gráfico é tranquilo...

Posição vs Ângulo



[Graphical perception:
Theory, Experimentation,
and Applications to the
Development of Graphical
Models](#)

Gráfico é tranquilo...

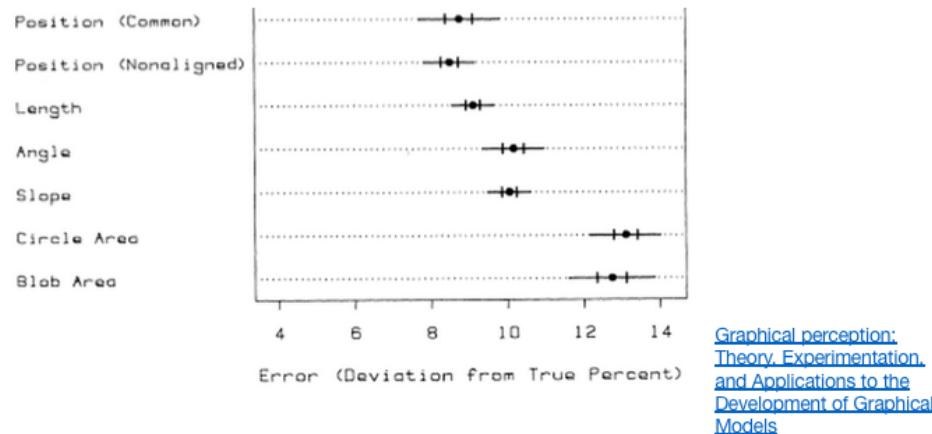


Gráfico é tranquilo...

Resumindo

- Use escalas comuns/padronizadas
- Use comparações de posição
- NÃO aos gráficos de pizza!
- NÃO aos gráficos em 3D



[Graphical perception: Theory, Experimentation, and Applications to the Development of Graphical Models](#)

[Graphical Perception and Graphical Methods for Analyzing Scientific Data](#)

Qual o gráfico correto para meus dados?

Qual gráfico
é o correto para você?

Conte histórias impactantes com dados

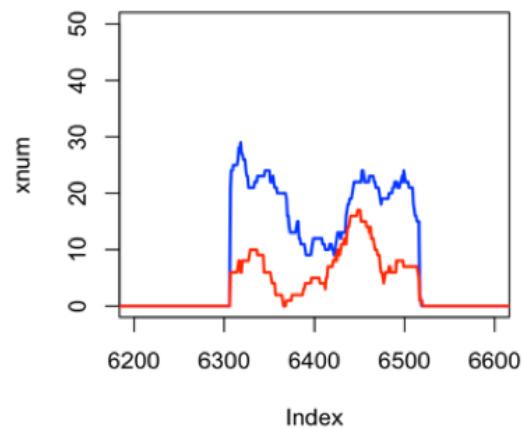


https://www.tableau.com/sites/default/files/whitepapers/which_chart_0512_pt.pdf

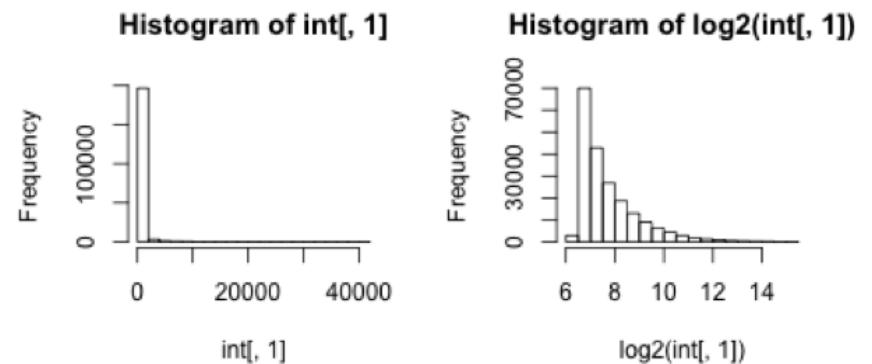
Análise Exploratória

Gráficos Exploratórios ≠ Gráficos Expositivos

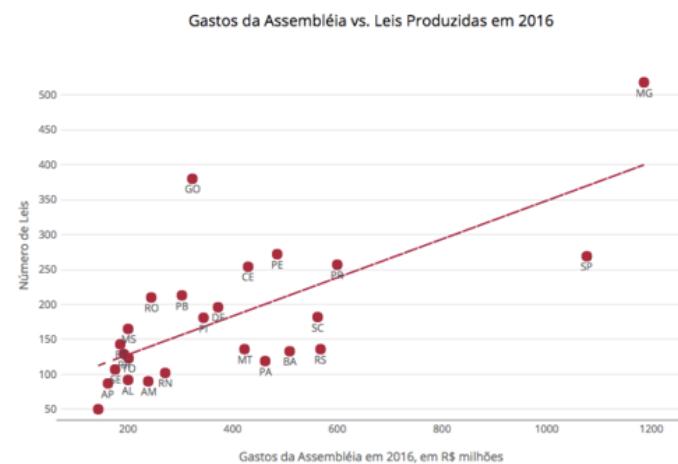
Análise Exploratória



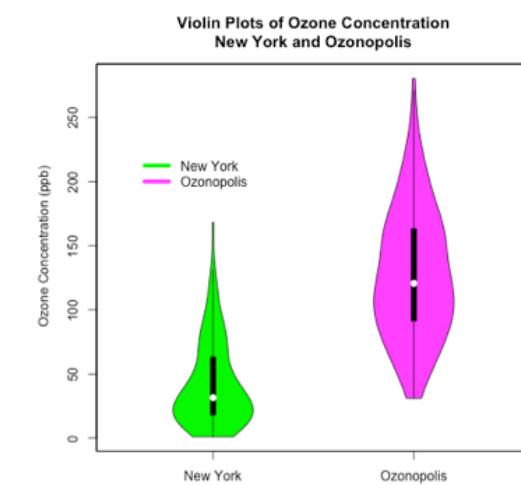
Análise Exploratória



Análise Exploratória



Análise Exploratória

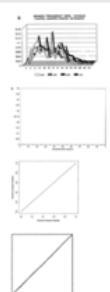


Análise Exploratória

The top ten worst graphs

With apologies to the authors, we provide the following list of the top ten worst graphs in the scientific literature. As these examples indicate, good scientists can make mistakes.

1. Roeder K (1994) DNA fingerprinting: A review of the controversy (with discussion). *Statistical Science* 9:222-278, Figure 4
[The article | The figure | Discussion]
2. Wittke-Thompson JK, Pluzhnikov A, Cox NJ (2005) Rational inferences about departures from Hardy-Weinberg equilibrium. *American Journal of Human Genetics* 76:967-986, Figure 1
[The article | Fig.1AB | Fig.1CD | Discussion]
3. Epstein MP, Satten GA (2003) Inference on haplotype effects in case-control studies using unphased genotype data. *American Journal of Human Genetics* 73:1316-1329, Figure 1
[The article | The figure | Discussion]
4. Mykland P, Tierney L, Yu B (1995) Regeneration in Markov chain samplers. *Journal of the American Statistical Association* 90:233-241, Figure 1
[The article | The figure | Discussion]



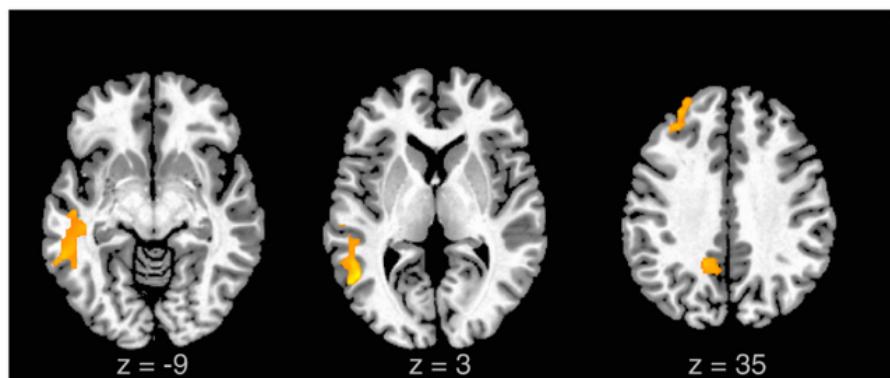
https://www.biostat.wisc.edu/~kbroman/topten_worstgraphs/

Análise Exploratória



<http://www.facebook.com/notes/facebook-engineering/visualizing-friendships/469716398919>

Análise Exploratória



[Liu et al. \(2012\) Scientific Reports](#)

Análise Exploratória – Eficiência Pregão

Em **85% dos órgãos** e mais de **30% dos pregões**:

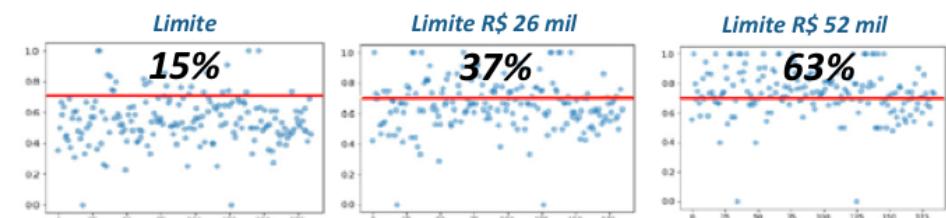
Custo de Pessoal > Benefício do Pregão

Nota Técnica
ao MP



EXEMPLO:

- Pregão de R\$ 20 mil → Gera economia média de 16% = R\$ 3,2 mil
- Economia equivale a 9 dias de trabalho (servidor que recebe R\$ 10 mil/mês)



Análise Exploratória – Eficiência Pregão

Nota Técnica
ao MP



Junho de 2018:
<http://www.planejamento.gov.br/noticias/s/decreto-atualiza-valores-para-licitacoes-e-contratos>

NOVOS VALORES LIMITE PARA AQUISIÇÕES PÚBLICAS POR MEIO DE LICITAÇÃO (alteração na Lei nº 8.666/1993)		
	CONVITE	TOMADA DE PREÇOS
OBRAS E SERVIÇOS DE ENGENHARIA	ANTES: Até R\$ 150 mil ↓ AGORA: Até R\$ 330 mil	ANTES: Até R\$ 1,5 milhão ↓ AGORA: Até R\$ 3,3 milhões
DEMAIS LICITAÇÕES (COMPRAIS E SERVIÇOS, EXCLUINDO-SE OBRAS E SERVIÇOS DE ENGENHARIA)	ANTES: Até R\$ 80 mil ↓ AGORA: Até R\$ 176 mil	ANTES: Até R\$ 650 mil ↓ AGORA: Até R\$ 1,43 milhão
		ANTES: Acima de R\$ 1,5 milhão ↓ AGORA: Acima de 3,3 milhões
		ANTES: Acima de R\$ 650 mil ↓ AGORA: Acima de R\$ 1,43 milhão

Contratações por meio de dispensa de licitação:
→ Valor máximo foi de R\$ 8 mil para R\$ 17,6 mil

Passagens aéreas

Estudo comparativo sobre o processo de aquisição de passagens aéreas antes e depois da adoção do novo modelo de compra direta pelo Governo Federal

Economia de R\$ 3,1 milhões entre janeiro e junho de 2016

- Preços
- Antecedência do processo
- Prazo de emissão dos bilhetes

Comparação Vertical

Comparação Horizontal

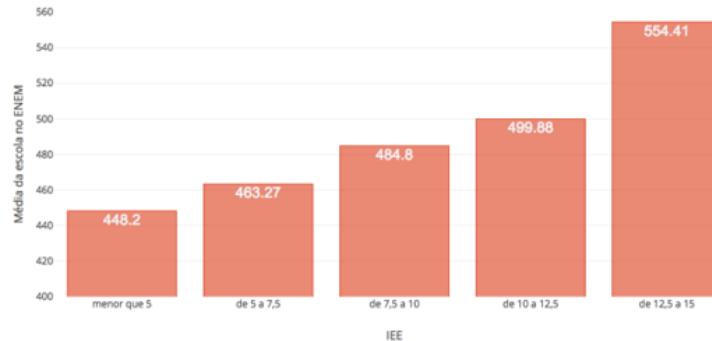
<http://www.cgu.gov.br/noticias/2016/09/governo-federal-economiza-r-3-1-mi-com-novo-modelo-de-compras-de-bilhetes-aereos>

http://www.cgu.gov.br/assuntos/informacoes-estrategicas/observatorio-da-despesa-publica/relatorios/relatorio_aquisicao_diarias-1.pdf

Análise Exploratória – ENEM

RELAÇÃO ENTRE ESTRUTURA ESCOLAR E DESEMPENHO NO ENEM

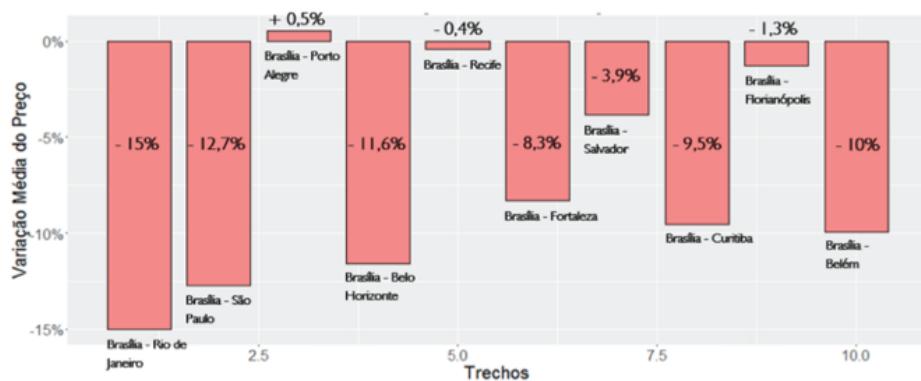
Índice de Estrutura da Escola X Média da Escola no ENEM



<https://leosallesblog.wordpress.com/2018/02/03/escola-ruim-aluno-ruim-entendendo-a-relacao-entre-estrutura-escolar-e-desempenho-no-enem/>

Passagens aéreas

Gráfico 1. Variação Média do Preço – Pareto 50



Passagens aéreas



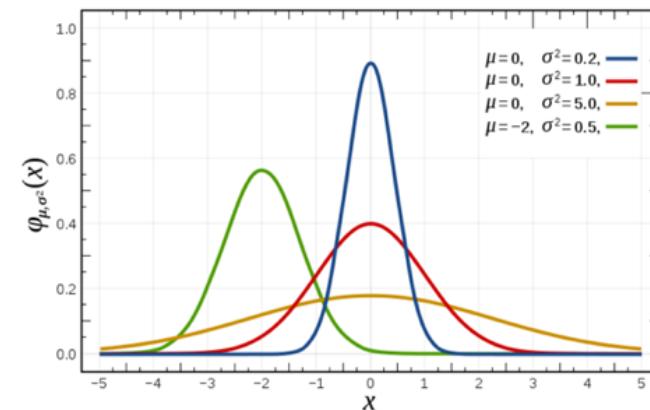
Tabela 3. Média dos preços praticados em 2016

Trecho	Compra Agenciada	Compra Direta	Variação	Teste de hipótese
BRASÍLIA/DF <-> RIO DE JANEIRO/RJ	R\$ 581,45	R\$ 494,06	-15,03%	Redução
BRASÍLIA/DF <-> SÃO PAULO/SP	R\$ 538,92	R\$ 470,34	-12,73%	Redução
BRASÍLIA/DF <-> PORTO ALEGRE/RS	R\$ 1.010,08	R\$ 1.015,64	0,55%	Indefinido
BRASÍLIA/DF <-> BELO HORIZONTE/MG	R\$ 459,75	R\$ 406,37	-11,61%	Redução
BRASÍLIA/DF <-> RECIFE/PE	R\$ 444,29	R\$ 442,40	-0,42%	Indefinido

Precisa disso aqui?
O que isso significa?

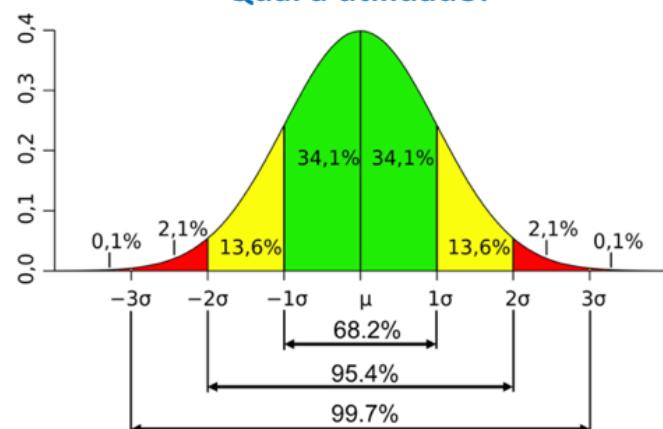
Distribuição de Probabilidade – Normal

O que é?



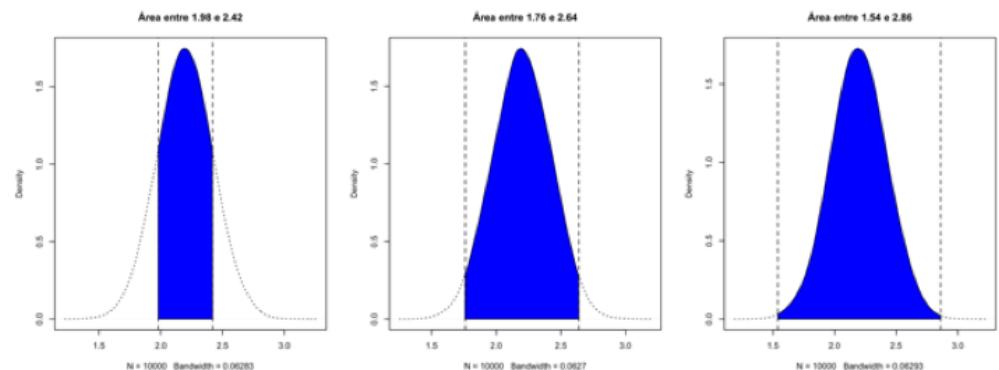
Distribuição de Probabilidade – Normal

Qual a utilidade?



Distribuição de Probabilidade – Normal

Então, com μ e σ estimados já temos..



Teste de Hipótese

Partimos de:

H0: Hipótese Nula

Situação assumida

Distribuição conhecida



H1: Hipótese Alternativa

O que se quer testar

Situação que advém da rejeição de H0

Nível de significância (α)

Normalmente 10%, 5% ou 1%

Chance de termos concluído errado pela rejeição da H0

Teste de Hipótese

Executando o teste:

- 1) Obtemos valor de teste em uma amostra
- 2) Assumimos hipótese nula como verdade, calculamos:
→ Probabilidade do valor em amostra estar na distribuição
→ Tal probabilidade é chamada de P-Valor
- 3) Baseado no resultado, rejeitamos ou não a situação assumida



P-valor < α

Rejeita-se a hipótese nula

P-valor > α

Nada posso afirmar

Teste de Hipótese



Partimos de:

H0: Hipótese Nula

Preço médio pago pelo GF por 20L de água mineral é o de mercado

Preço médio de mercado é R\$ 25,00



H1: Hipótese Alternativa

Preço médio pago pelo GF por 20L de água mineral é maior que o de mercado

Nível de significância (α)

Arbitrariamente 5%

Teste de Hipótese



Executando o teste:



[Amostra GF] Preço médio: R\$ 31,70

Assumimos hipótese nula como verdade:

Ou seja, Preço médio GF = Preço médio do Mercado (Distribuições)

Calculamos P-valor:

→ Probabilidade do valor em amostra estar na distribuição assumida

Lembrando que selecionamos $\alpha = 0.05$

P-valor = 0.03 < α

Rejeita-se a hipótese nula

Preço GF > Preço de mercado

P-valor = 0.09 > α

Nada posso afirmar

Não posso aceitar H0!

Passagens aéreas



Tabela 3. Média dos preços praticados em 2016

Trecho	Compra Agenciada	Compra Direta	Variação	Teste de hipótese
BRASÍLIA/DF <-> RIO DE JANEIRO/RJ	R\$ 581,45	R\$ 494,06	-15,03%	Redução
BRASÍLIA/DF <-> SÃO PAULO/SP	R\$ 538,92	R\$ 470,34	-12,73%	Redução
BRASÍLIA/DF <-> PORTO ALEGRE/RS	R\$ 1.010,08	R\$ 1.015,64	0,55%	Indefinido
BRASÍLIA/DF <-> BELO HORIZONTE/MG	R\$ 459,75	R\$ 406,37	-11,61%	Redução
BRASÍLIA/DF <-> RECIFE/PE	R\$ 444,29	R\$ 442,40	-0,42%	Indefinido

O que aconteceu aqui??

Análise Inferencial

Intervalos de confiança

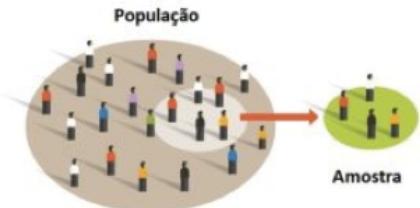
$$\begin{array}{lll} 34 + 2\% & 29 + 2\% & 19 + 2\% \\ 34 - 2\% & 29 - 2\% & 19 - 2\% \end{array}$$



Análise Inferencial

OBJETIVO: Usar amostra para dizer algo sobre população

- Estimar uma quantidade de interesse
- Estimar a incerteza relacionada com a estimativa da quantidade de interesse



O que usaremos?

Teste de Hipótese

Intervalos de confiança

Modelos estatísticos (regressão linear, logística..)

Análise Inferencial

Effect of Air Pollution Control on Life Expectancy in the United States: An Analysis of 545 U.S. Counties for the Period from 2000 to 2007

Correia, Andrew W.^a; Pope, C. Arden III^b; Dockery, Douglas W.^c; Wang, Yun^d; Ezzati, Majid^d; Dominici, Francesca^e

Epidemiology: January 2013 - Volume 24 - Issue 1 - p 23-31
doi: 10.1097/EDE.0b013e3182770237
Air Pollution

Results: A decrease of $10 \mu\text{g}/\text{m}^3$ in the concentration of $\text{PM}_{2.5}$ was associated with an increase in mean life expectancy of 0.35 years ($SD = 0.16$ years, $P = 0.033$). This association was stronger in more urban and densely populated counties.

https://journals.lww.com/epidem/Fulltext/2013/01000/Effect_of_Air_Pollution_Control_on_Life_Expectancy.4.aspx

Análise Preditiva

OBJETIVO: Usar dados de uns objetos para prever valores de outros

Aprendizagem de Máquina:

"A máquina aprende com a experiência, sem ser explicitamente programada para isso."

Exemplo:



Machine Learning



"Todos os modelos estão errados, porém alguns são úteis"

– George E. P. Box

Modelos são uma **aproximação** de algo que observamos

Simplificação da realidade

→ Mas algumas aproximações podem ser **úteis** para nós

Exemplos:



Machine Learning

Risco de Crédito = ALTO/... /BAIXO



Wearables (Vestíveis) = Tipo de movimento



Netflix/Amazon = Recomendação



Walmart = Regras de Associação



Machine Learning

Dados melhores geralmente vencem modelos maiores

Melhores dados >> melhores algoritmos >> melhores parâmetros

Não existe bala de prata



Lembre-se de Occam's Razor

Exemplo: Netflix prize

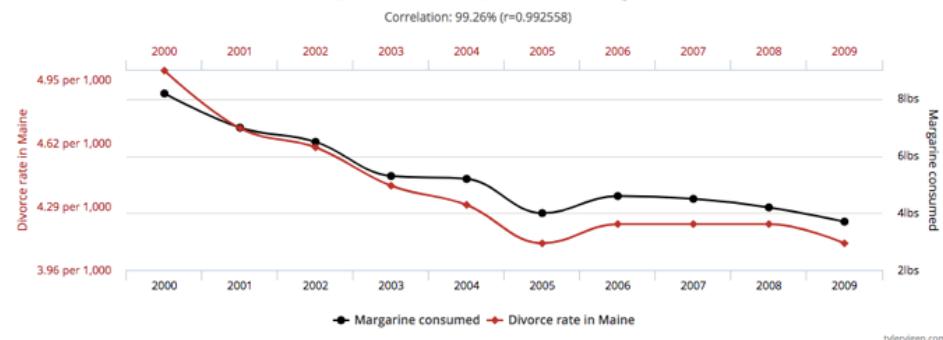


Se X prevê Y isso não significa que X causa Y

Correlação não implica Causalidade



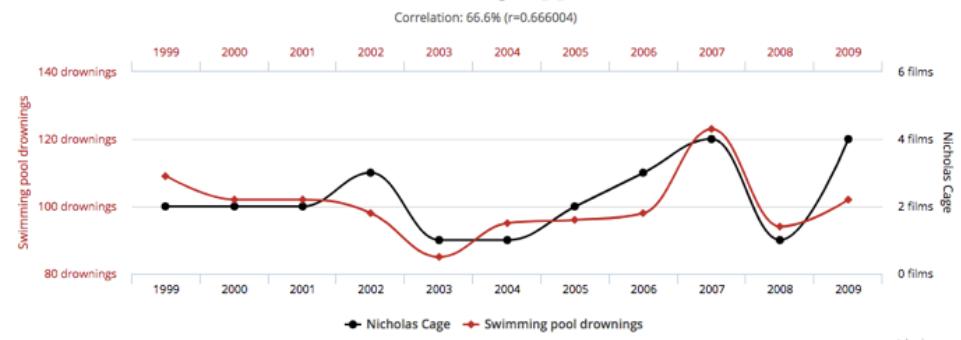
Divorce rate in Maine
correlates with
Per capita consumption of margarine



Correlação não implica Causalidade



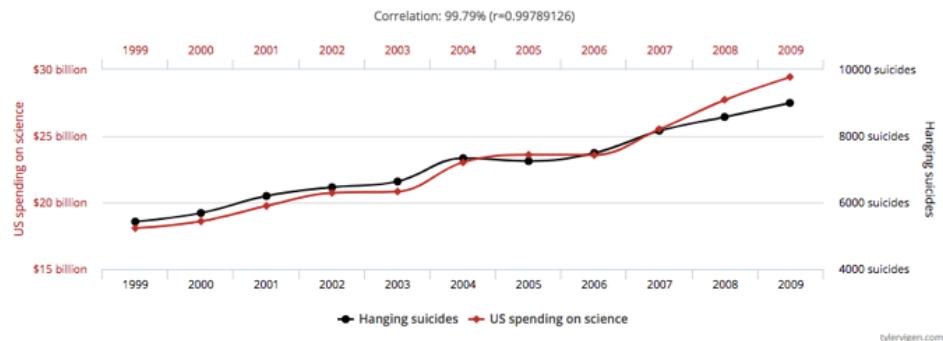
Number of people who drowned by falling into a pool
correlates with
Films Nicolas Cage appeared in



Correlação não implica Causalidade



US spending on science, space, and technology
correlates with
Suicides by hanging, strangulation and suffocation



Correlação não implica Causalidade



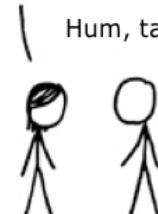
Eu costumava achar
que correlação
implicava causalidade



Então eu assisti a
uma aula de
estatística, agora
não acho mais.



Parece que a aula
ajudou.
Hum, talvez.



Análise Preditiva



Análise Preditiva

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did



Kashmire Hill, FORBES STAFF

Every time you go shopping, you share intimate details about your consumption patterns with retailers. And many of those retailers are studying those details to figure out what you like, what you need, and which coupons are most likely to make you happy. Target, for example, has figured out how to data-mine its way into your womb, to figure out whether you have a baby on the way long before you need to start buying diapers.



Target has got you in its aim

<http://www.forbes.com/sites/kashmirehill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/>

Análise Inferencial vs Preditiva

- Por exemplo, em um cenário imobiliário, pode-se procurar **relacionar os preços** das casas a **variáveis** como taxa de criminalidade, zoneamento, distância de um rio, qualidade do ar, escolas, nível de renda da comunidade, tamanho das casas e assim por diante.
- Nesse caso, pode-se estar interessado em **como as variáveis de entrada individuais afetam os preços** – ou seja, quanto mais uma casa valerá se tiver uma vista do rio? Esse é um problema de **inferência**.
- Por outro lado, pode-se simplesmente estar interessado na **previsão do preço de uma casa, dadas as suas características**: esta casa está subvalorizada ou supervalorizada? Este é um problema de **previsão**.

[An Introduction to Statistical Learning](#)

Análise Causal e Análise Mecanicista

Análise Causal

Objetivo: Descobrir o que acontece com uma variável quando você altera outra

- Geralmente uso de estudos randomizados
- Efeito médio, pode não ser aplicável a um indivíduo
- **Exemplo:** Teste A/B



Análise Mecanicista

Objetivo: Entender as exatas mudanças em variáveis que levam a mudanças em outras variáveis para objetos individuais

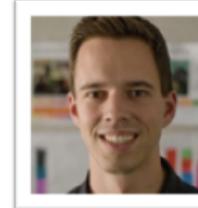
- Normalmente conjunto determinístico de equações (física/engenharia)
- **Exemplo:** Análise de dados de pavimentos em rodovias

Análise de dados: Na prática

Projetos de Análise de Dados

“Pergunte a si mesmo, que problema você já resolveu, que valeu a pena resolver, onde você conhecia todas as informações fornecidas com antecedência?

Onde você **não** tinha um excesso de informações e precisava filtrá-las
ou
onde você já tinha informações suficientes e **não** precisava encontrar mais nenhuma?”



Dan Myer, Educador

Projetos de Mineração de Dados

CRISP-DM

- Cross Industry Process Model for Data Mining

- Hierárquico:

- Fases
- Tarefas genéricas
- Tarefas especializadas
- Registro de ações



Projetos de Mineração de Dados

CRISP-DM

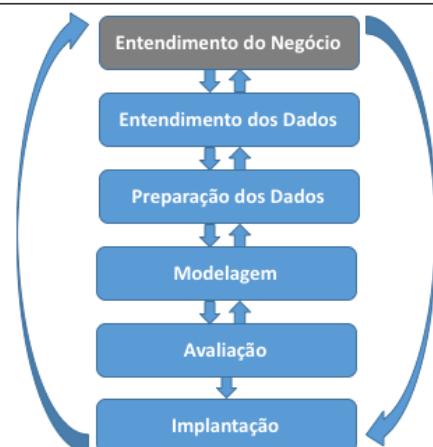
Entendimento do Negócio

NEGÓCIO

- Requerimentos do projeto
- Objetivos
- Perspectiva do domínio
- Conhecimento prévio

- Plano do projeto:

- Passos a serem executados no resto do projeto e a definição do problema.

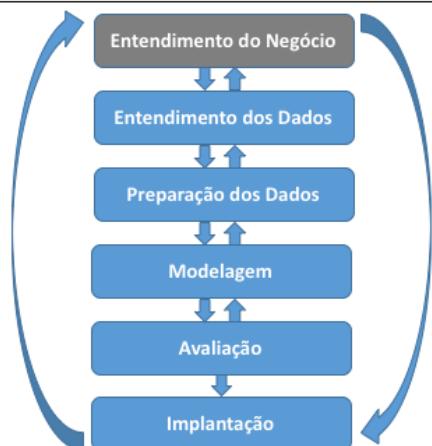


Projetos de Mineração de Dados

CRISP-DM

Entendimento do Negócio

8 passos para aproximar negócio e ciência de dados



Projetos de Mineração de Dados

CRISP-DM

Preparação dos Dados

QUALIDADE

Integração de dados

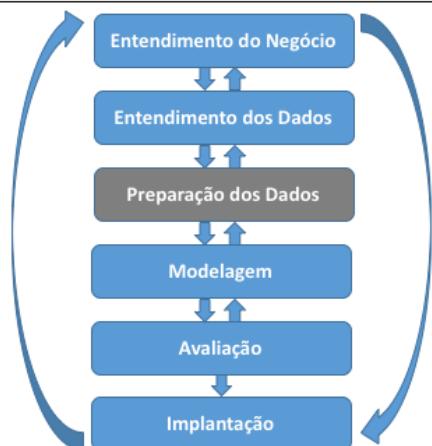
Seleção de atributos

Limpeza e padronização

Criação de atributos derivados

Tratamento de valores faltantes

Análise de anomalias (outliers)



Projetos de Mineração de Dados

CRISP-DM

Entendimento dos Dados

SIGNIFICADO

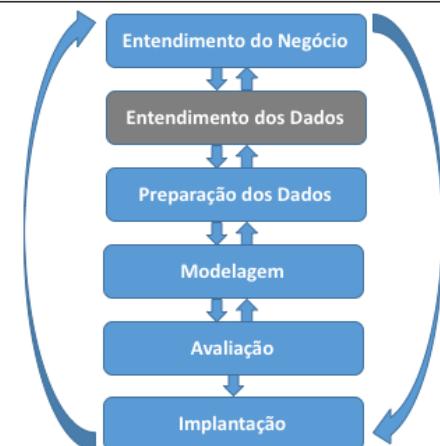
Seleção dos dados que vou usar

Descrição e Formato dos dados

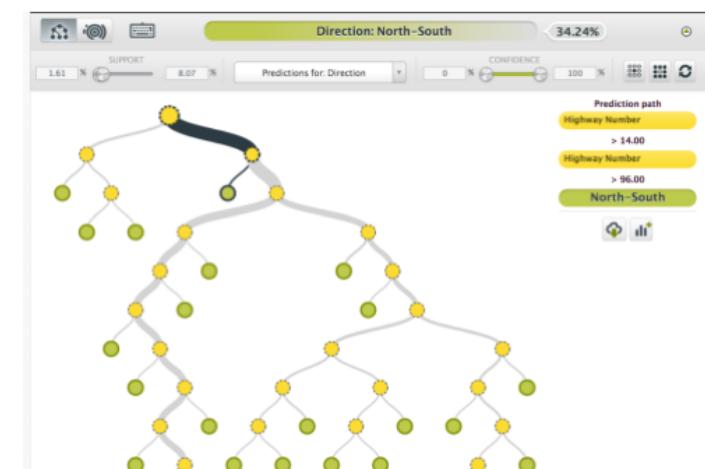
Identificar problemas

Relacionamentos entre atributos

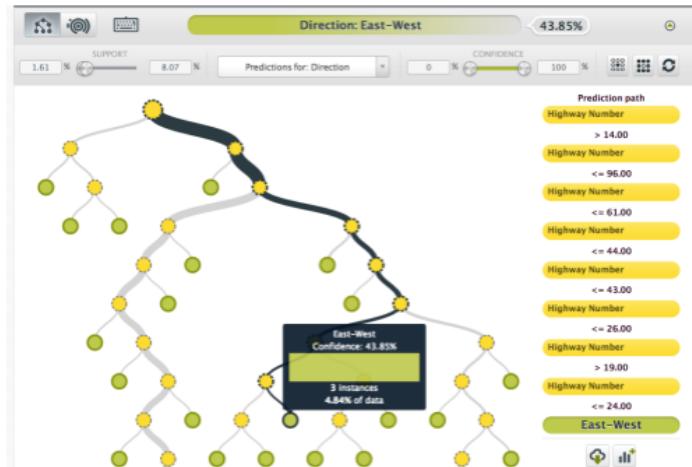
Identificação de agrupamentos



Feature Engineering



Feature Engineering



Feature Engineering



Feature Engineering

Titanic dataset

PassengerId	Survived	Pclass	Name				
1	0	3	Braund, Mr. Owen Harris				
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...)				
3	1	3	Heikkinen, Miss. Laina				
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)				
5	0	3	Allen, Mr. William Henry				
Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
male	22.0	1	0	A/5 21171	7.2500	Nan	S
female	38.0	1	0	PC 17599	71.2833	C85	C
female	26.0	0	0	STON/O2. 3101282	7.9250	Nan	S
female	35.0	1	0	113803	53.1000	C123	S
male	35.0	0	0	373450	8.0500	Nan	S

Feature Engineering

Titanic dataset

Name	title
Braund, Mr. Owen Harris	Mr
Cumings, Mrs. John Bradley (Florence Briggs Th...)	Mrs
Heikkinen, Miss. Laina	Miss
Futrelle, Mrs. Jacques Heath (Lily May Peel)	Mrs
Allen, Mr. William Henry	Mr

Feature Engineering = Entend. + Prepar.

O QUE PODEMOS DERIVAR DE:

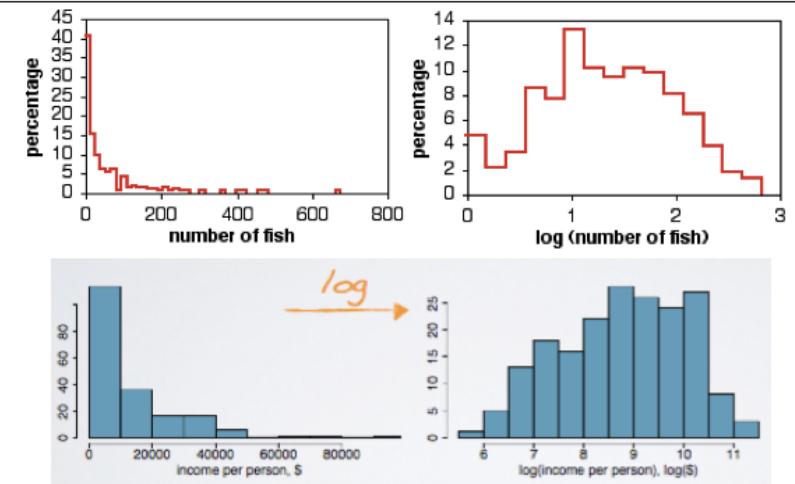
- Data de início de filiação + Data de término?
- Data de início em cargo público?
- Salário de servidor por vários meses?
- Campo Órgão como "49000 – Secretaria de XYZ"?



TRANSFORMAÇÃO:

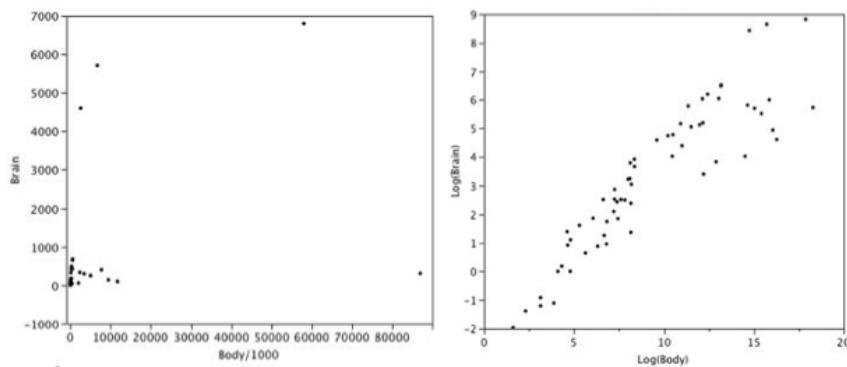
- Binarizar: Dummy Variables, One-hot encoding
- Log
- Normalizar
- Discretizar

Feature Engineering = Entend. + Prepar.

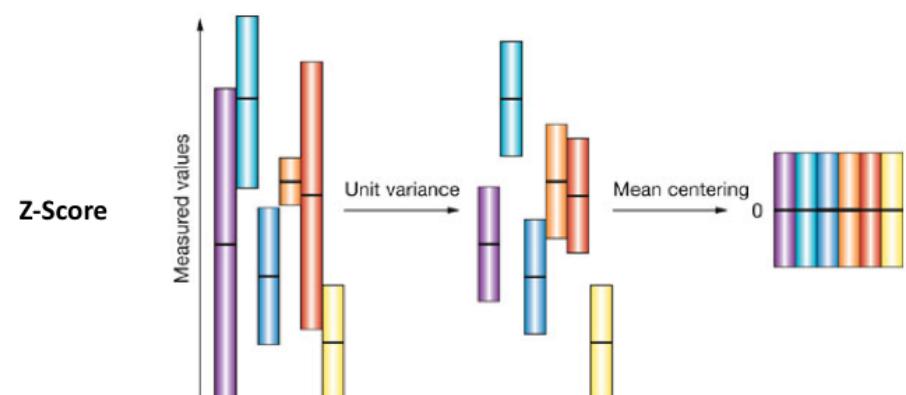


Feature Engineering = Entend. + Prepar.

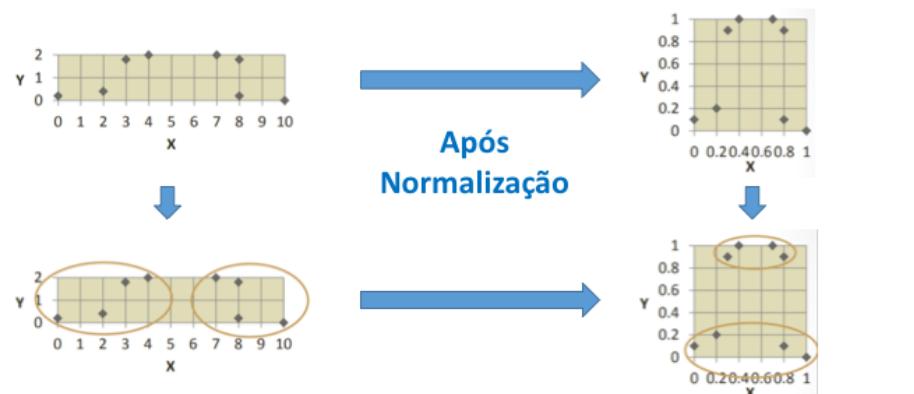
LOG



Normalização



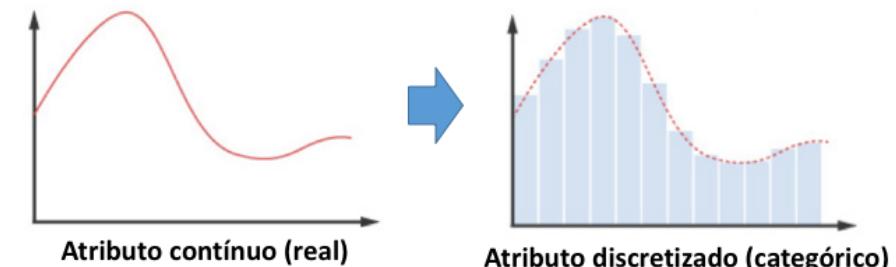
Normalização



Discretização

Diminuir variabilidade → Simplifica representação

Exemplo: Transformar “Altura” em “Faixas de Altura”



Limpeza e Padronização

- **Sexo:** “M”, “F”, “Masc”, “Fem”, “Masculino”
- **Data inicialmente texto:**
 - **Formato:** “2010-03-22”, “22/03/2010”
 - **Erro “bruto”:** “2010-13-22”, “13/13/2010”
 - **Erro “lógico”:** “02/12/2018” é um Domingo
- **UF:** “Santa Catarina”, “Santa Catairna”
- **Idade:** 22, 28, 30, -20, 11, 188
- **Possui casa própria:** “Sim”, “Não”, “VAZIO”, “ ”
- **Filtros:** Quero analisar aposentados, pensionistas?
- **Como trato Valores faltantes?**
- **Como encontro Outliers e como tratá-los?**
- **Como extraio valor de textos livres/maiores?**



Outros cenários de análise



Falta de dado ou realidade?

Existe dado que separe tão bem assim as classes?

→ Analisar **Variância POR CLASSE** das variáveis
Próxima de zero

sg_partido	nm_dasse
PT	1
PSDB	0
PDT	0
PSOL	0
PMDB	0
PV	0

Outros cenários de análise



2013		2014		2015		2016	
Mês	Qtd.	Mês	Qtd.	Mês	Qtd.	Mês	Qtd.
1	231	1	123	1	72	1	231
2	123	2	234	2	34	2	723
3	876	3	734	3	436	3	955
4	14	4	22	4	325	4	16
5	34	5	63	5	234	5	124
6	24	6	45	6	33	6	546
7	656	7	0	7	64	7	315
8	424	8	0	8	144	8	254
9	243	9	0	9	134	9	164
10	34	10	0	10	37	10	69
11	66	11	0	11	99	11	17
12	46	12	94	12	83	12	25

Posso analisar os 4 anos juntos?
E comparar médias dos 4?

O que houve em 2014/2?
Minha amostra é confiável?

Estratégias para valores faltantes



Faltas aleatórias: **não** há padrão nos dados ausentes em nenhuma variável.

Faltas NÃO aleatórias: **há** padrão nos dados ausentes que afetam variáveis.

Média, Mediana, Moda: Quando forem Faltas aleatórias

Constante: Quando forem Faltas NÃO aleatórias afetando variável **dependente**

Substituir por “chute”: Quando possível inferir valor (use com muita moderação!)

Deletar coluna(s) : Quando houver excesso de faltantes

Incerteza na aleatoriedade: Modelos probabilísticos de imputação

Deletar?

Só se amostra for grande o suficiente e for **Falta aleatória...** senão pode estar excluindo algum grupo importante.

Outliers

Premissa:

Há consideravelmente mais observações “normais” do que observações “anormais” (outliers/anomalias) nos dados



Passo a passo:

Construir um perfil do comportamento “normal”

Anomalias são dados cujas características **diferem significativamente** do perfil normal

O que fazer após encontrar outliers (se não for objetivo preditivo):

Eliminar

Separar para análise diferenciada

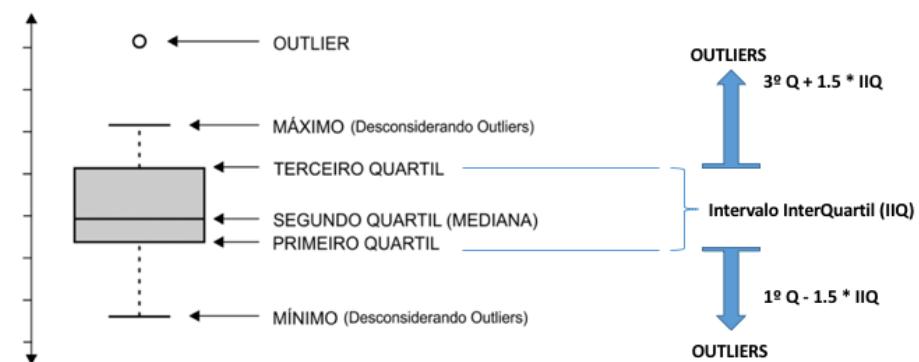
Usar peso diferente para outlier

Usar score de “outlier” (idade como novo atributo do modelo)

Relatar

Detecção de Outliers

Gráfico ou baseado em estatísticas



Detecção de Outliers

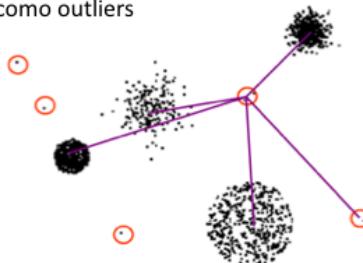
Baseado em distâncias

Agrupa os dados em grupos de diferentes densidades

Define os pontos em pequenos clusters como outliers candidatos

Calcula a distância entre os outliers candidatos e os clusters não candidatos.

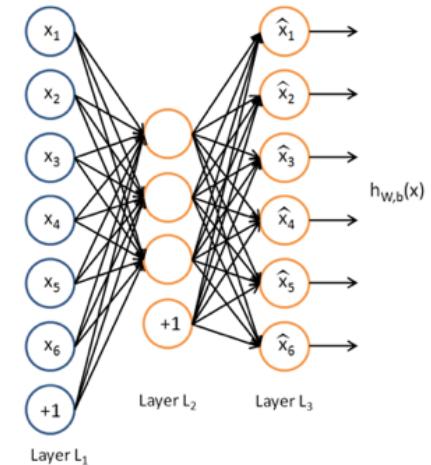
Se os pontos candidatos estiverem longe de todos os outros clusters não candidatos, eles serão confirmados como outliers



Detecção de Outliers

Baseado em modelos

Deep Learning: Autoencoder



Como extrair valor de textos?

Mineração de Texto



Representação

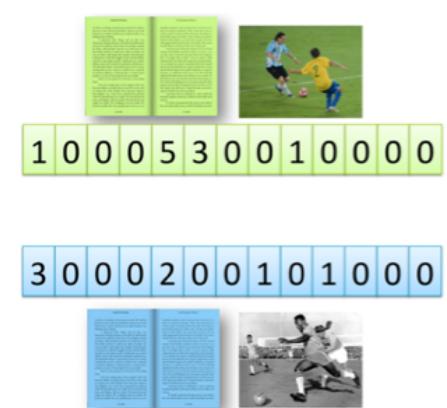


Mineração de Texto

Medindo Similaridade

$$= 1 * 3 + 5 * 2$$

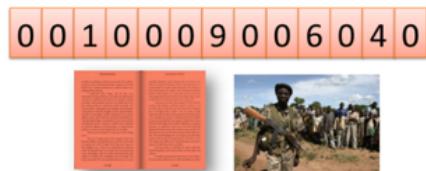
$$= 13$$



Mineração de Texto

Medindo Similaridade

= 0



Mineração de Texto

Medindo Similaridade

2º problema:

Palavras comuns (jogador, gol, campo)
→ Dominam as raras (Messi, voleio)

Solução:

Usar palavras importantes!

O que seria uma palavra importante?

Aparece muito no documento (localmente comum/frequente)
Aparece raramente em corpus (raro globalmente)



Mineração de Texto

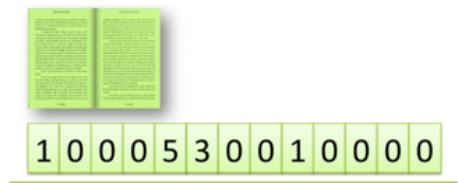
Medindo Similaridade

1º problema:

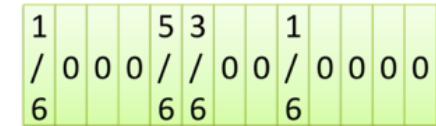
Textos grandes
→ Números grandes

Solução:

Normalizar!



$$\sqrt{(1^2 + 5^2 + 3^2 + 1^2)}$$



Mineração de Texto

Mineração de Texto

Medindo Similaridade

Usar palavras importantes!

O que seria uma palavra importante?

Aparece muito em documento (localmente frequente)
Aparece raramente em corpus (raro globalmente)

TF-IDF

Term frequency * Inverse Document Frequency
Frequência termo * Inverso Frequência Documento
= Trade-off

Em TF-IDF geralmente aplica-se log

Muito em doc e corpus
grande/(1+grande) ~ log 1 ~ 0

Muito no doc, rara no corpus
grande/(1+pequeno) ~ grande

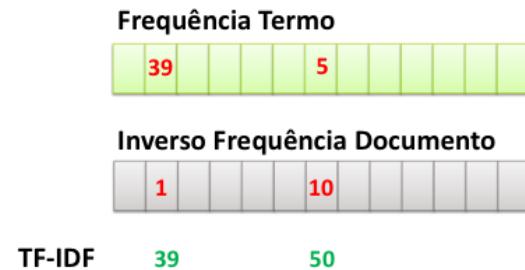
Mineração de Texto

TF-IDF

Term frequency * Inverse Document Frequency

Frequência termo * Inverso Frequência Documento

= Trade-off



Mineração de Texto: Sentimento

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Vivamus ut aliquam nisi. Sed vitae metus a nulla. Quisque pulvinar. Quisque pulvinar. Quisque pulvinar.

POSITIVAS	NEGATIVAS
bom	ruim
ótimo	péssimo
gostei	odeio
legal	chato
excelente	horrible

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Vivamus ut aliquam nisi. Sed vitae metus a nulla. Quisque pulvinar. Quisque pulvinar. Quisque pulvinar.

TERMO	SCORE
bom	+1
péssimo	-4
excelente	+4
chato	-0.5
ruim	-1

Nr. Positivas > Nr. Negativas

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Vivamus ut aliquam nisi. Sed vitae metus a nulla. Quisque pulvinar. Quisque pulvinar. Quisque pulvinar.

Soma Score > 0

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Vivamus ut aliquam nisi. Sed vitae metus a nulla. Quisque pulvinar. Quisque pulvinar. Quisque pulvinar.

Mineração de Texto

Operações usuais:

- Remoção de StopWords
- Padronização maiúscula/minúscula
- Stemming
- Negação
- Remoção de pontuação, números, espaços

Eu **não** gostei deste filme, mas a ...

Eu **não** NAO_gostei NAO_deste NAO_filme, mas a ...

EXCELENTE

Excelente
excelente

livro

livrinho

livreiro

livreco

participa

participo

participamos

Projetos de Mineração de Dados

CRISP-DM

Modelagem

OTIMIZAÇÃO

Função do tipo de dados (problema)

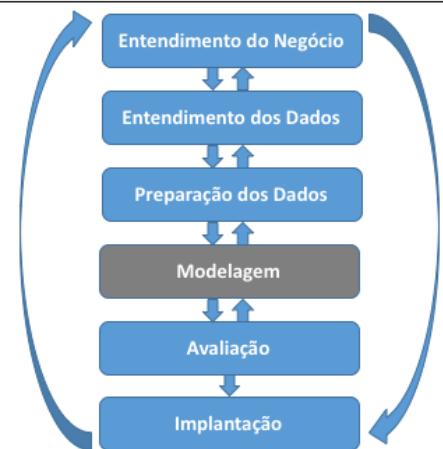
Quais algoritmos?

Quais parâmetros dos algoritmos?

- Plano de testes

Como avaliar os modelos gerados?

Divisão em treino/teste/validação



Projetos de Mineração de Dados

CRISP-DM

Avaliação

CRÍTICA

Desafie todas as etapas!

Pergunta

Fonte de dados

Processamento

Análise

Conclusões

Pense em análises alternativas



Projetos de Mineração de Dados

CRISP-DM

Implantação

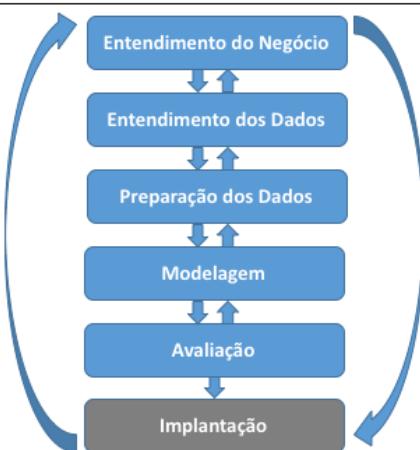
AÇÃO

Resultado incorporado à organização

Interpretar e sintetizar

- Plano de monitoração e manutenção

Prevenir uso incorreto dos resultados, ao longo do tempo



Projetos de Mineração de Dados

CRISP-DM

Avaliação

Desempenho nos dados de treino NÃO é bom indicador de desempenho em dados independentes

- Regressão:
 $(0 \leq R^2 \leq 1)$
- Classificação:
Acurácia, Precisão, F1, ...



Projetos de Mineração de Dados

Como alcanço meu público?

Modelo embutido em plataformas

Dashboard

E-mails e/ou mensagens periódicas

Relatório mensal

Relatório único

Uma página

Um gráfico

Um parágrafo



Projetos de Mineração de Dados

Relatório de Análise / Resumo

- Resumir as análises em uma história
 - 1) Começar com a pergunta
 - 2) Ordenar de acordo com a história, em vez de cronologicamente
- Não incluir todas as análises
- Incluir figuras “bonitas” que contribuam para a história



Jupyter Notebook

R Markdown

Projetos de Mineração de Dados

REPRODUCIBILIDADE

Código Reproduzível (organizado e comentado!)

Organização de pastas

Máquinas Virtuais / Contêineres Docker

Seed

Relatório de Análise



“Alguém sem familiaridade com o projeto deve conseguir ver os arquivos e entender com detalhes o que foi feito e o porquê”

Projetos de Mineração de Dados

Relatório de Análise / Resumo

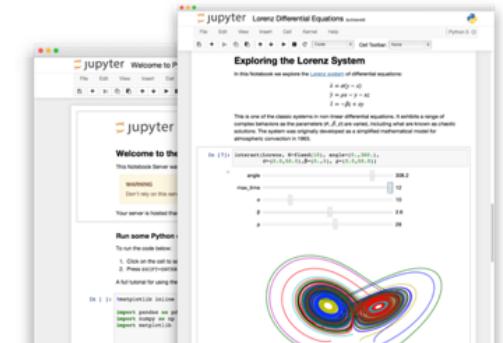
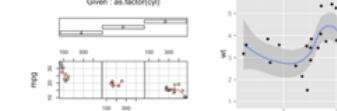
Motor Trend Car Road Tests

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the Knit button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

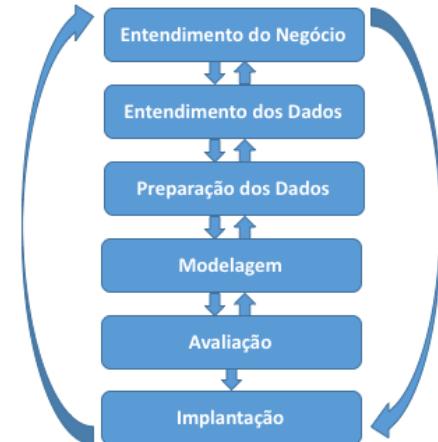
```
## # speed   dist
## 1 Min. ~12.0  2.00
## 2 1st Qu.~13.0  9.00
## 3 Median~15.0 26.00
## 4 3rd Qu.~19.0 43.00
## 5 Max.~29.0 120.00
```

You can also embed plots to help tell your story.



Projetos de Mineração de Dados

CRISP-DM remains the most popular methodology for analytics, data mining, and data science projects, with 43% share in latest KDnuggets Poll



CONCLUSÃO: Preparem-se!

Análise de Dados = Preparação + Entendimento de dados

80% do tempo nessas etapas 🍌🍌🍌

“Adoramos o caos porque adoramos produzir ordem.” – M.C.Escher

“Oh, sabe de uma coisa sobre o caos? É justo.” – Coringa

Curso foca nisso!!

