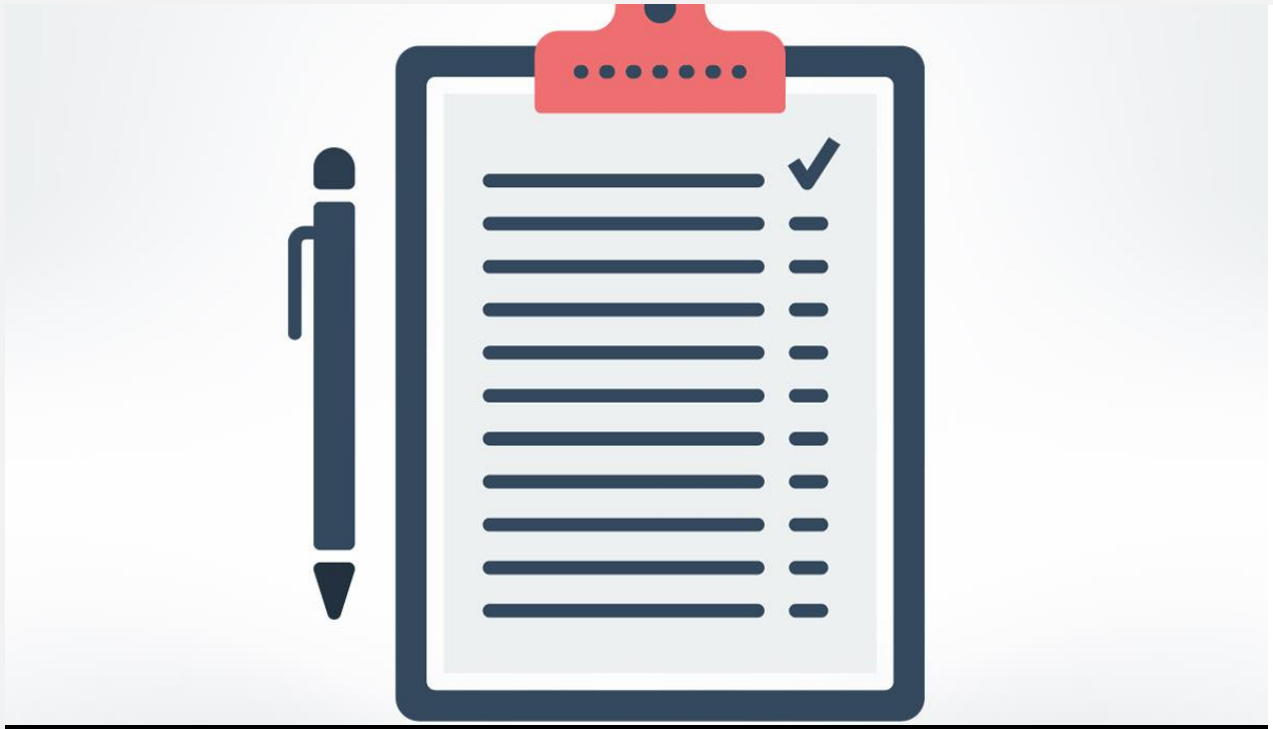
Mineração de Dados

Metodologia + Compreendendo Dados

Eduardo Silva – easilva91@gmail.com



Agenda

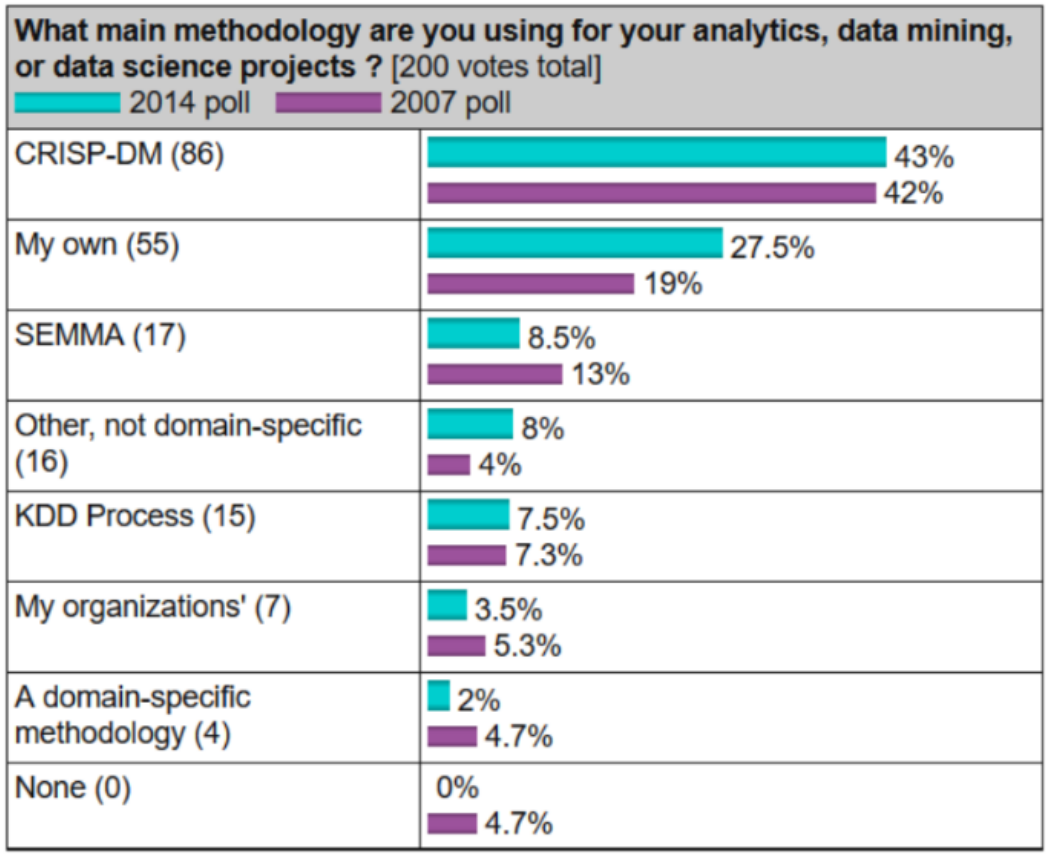
Mineração de Dados

1. Metodologia
   I. CRISP-DM
2. Compreensão de Dados
   I. Estatística resumida/básica
   II. Visualização de Dados
      I. Exemplos
   III. Analise Exploratória de Dados
   IV. Referências

Metodologias em Mineração de Dados

Mineração de dados assim como outros tipos de análise, tem diversas metodologias possíveis ou frameworks que podem ser seguidos, dentre eles, os mais comuns provavelmente são o CRISP-DM e o SEMMA.

- Cross-Industry Stadard Process for Data Mining (CRISP-DM)
- Sample, Explore, Modify, Model, and Assess (SEMMA)



https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

CRISP-DM

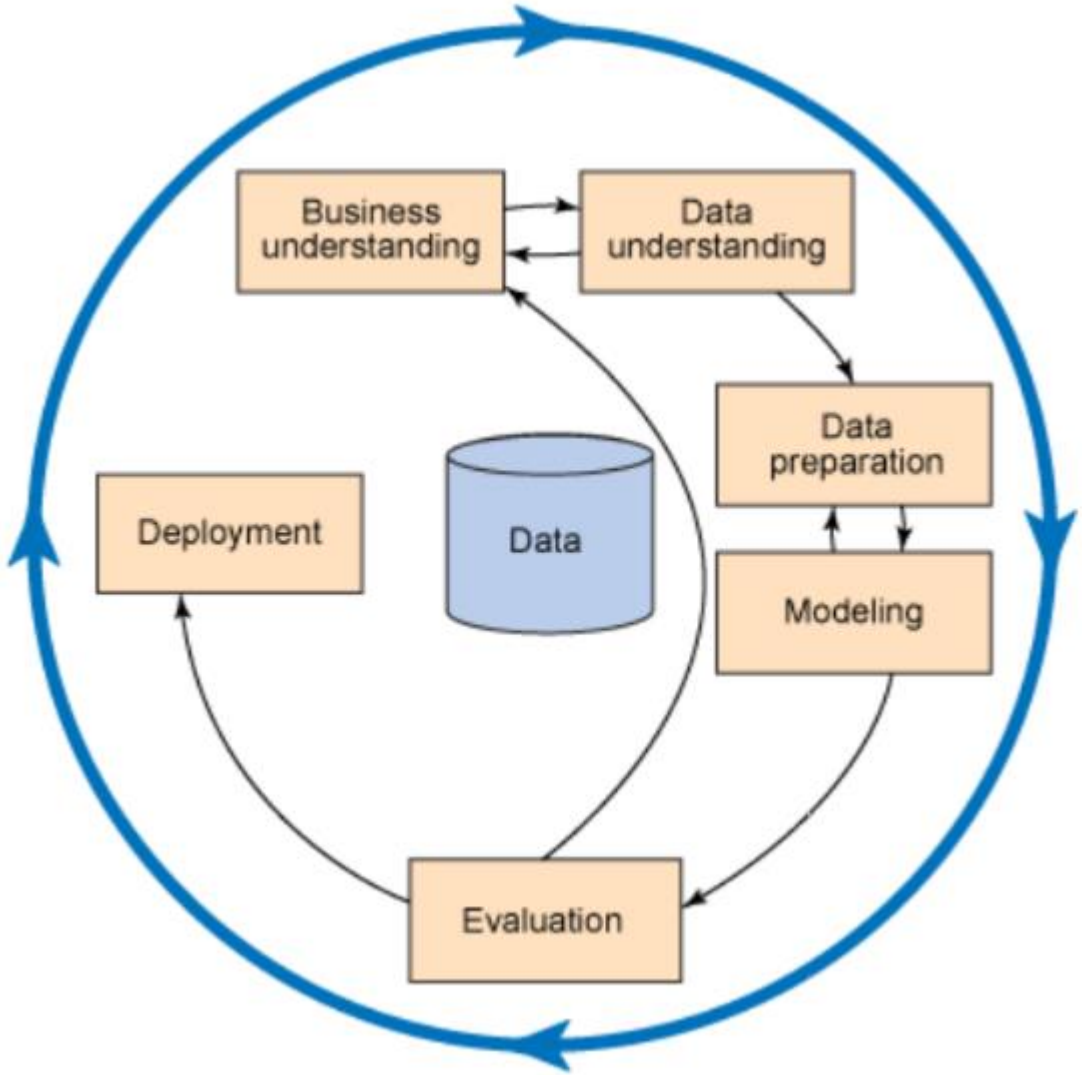
- Inicialmente lançado no fim de 1996 por três “veteranos” do mercado de mineração de dados.
  - Daimler Chrysler (Daimler-Benz), SPSS (ISL), NCR
- Desenvolvido e Refinado através de uma série de workshops (de 1997-1999)
- Mais de 300 organizações contribuíram para o modelo
- Publicação do CRISP-DM 1.0 (1999)
- Mais de 200 membros do CRISP-DM SIG pelo mundo.
  - DM Vendors – SPSS, NCR, IBM, SAS, SGI.
  - System Suppliers / consultaores – Cap Gemini, ICL Retail, Deloitte & Touche.
  - Usuários Finais – BT, ABB, Lloyds Bank, AirTouch.

CRISP-DM

- Não proprietário
- Aplicação Neutra
- Ferramenta Neutra
- Focado em problemas de Negócio
  - Assim como em analises técnicas
- Framework como guia
- Baseado em experiencia
  - Template para analises

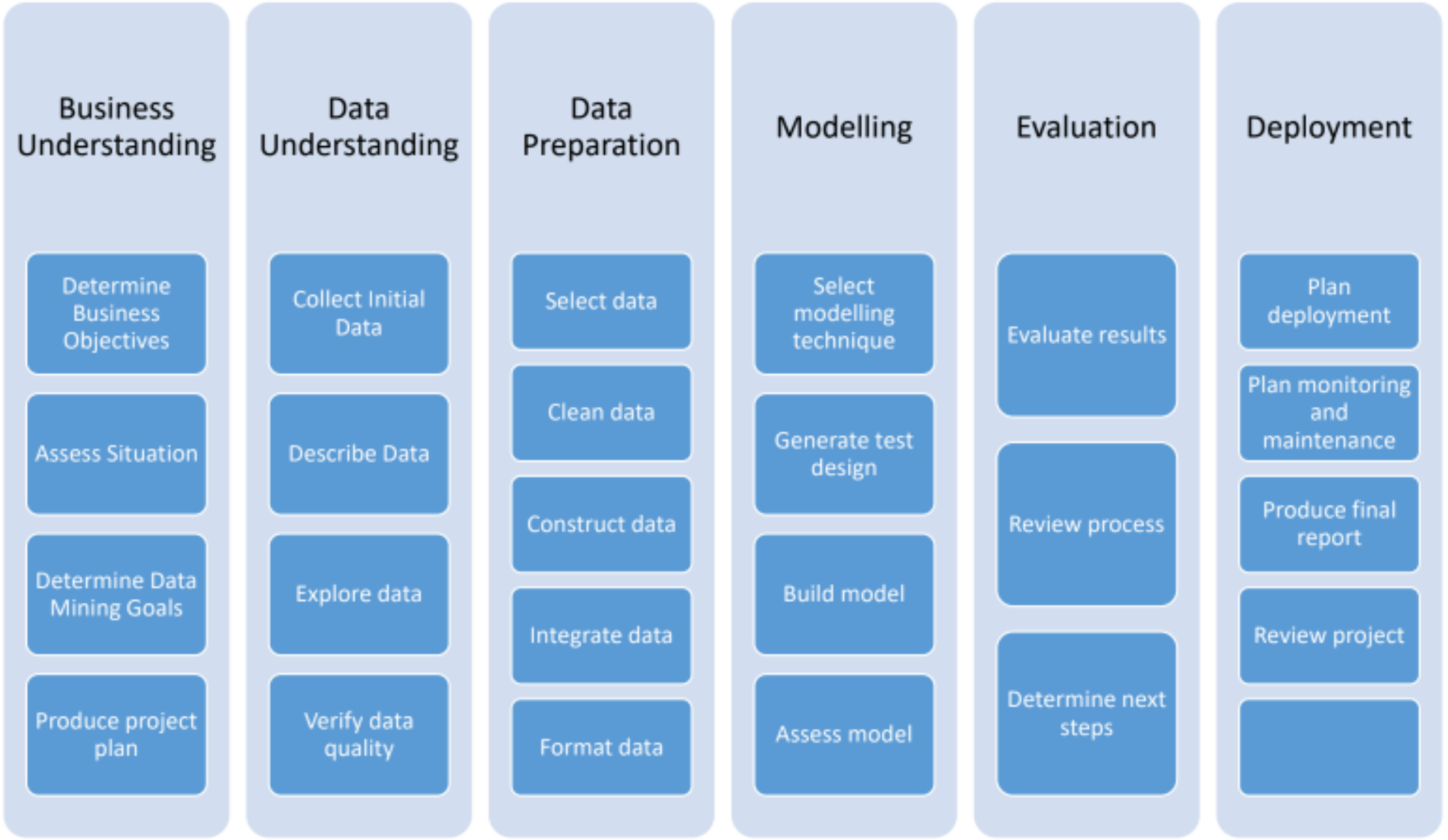
CRISP-DM

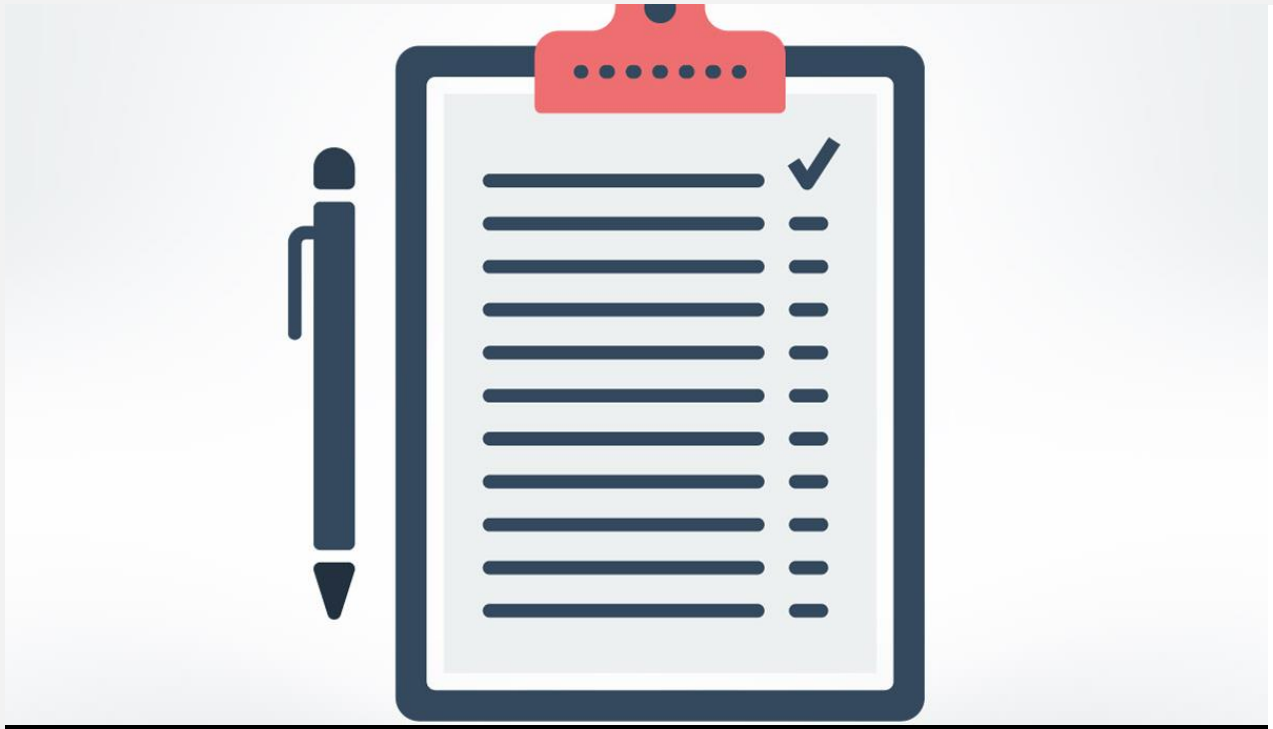
- Metodologia de Mineração de Dados
- Modelo de Processamento
- Para todos
- Fornece um modelo completo
- Ciclo de vida: 6 Fases



CRISP-DM

- Fases e Tarefas





Agenda

Mineração de Dados

1. Metodologia
2. Compreensão de Dados
   I. Estatística resumida/básica
   II. Visualização de Dados
      I. Exemplos
   III. Analise Exploratória de Dados
   IV. Referências

Compreensão de Dados

- Examina as principais características dos dados, incluindo
  - Quantos registros estão disponíveis
  - Quantas variáveis
  - Quantas variáveis de destino ou target variables (variável que deve ser o output em um algoritmo de ML)
- Comece por enumerar problemas com os dados, incluindo valores incorretos ou inválidos, valores em falta, distribuições inesperadas, e outliers (anomalias).
- Visualize os dados para ganhar ainda mais insights em relação as características do mesmo, principalmente aquelas mascaradas por estatística básica.

Tipos de Dados

Tipo de Atributo	Descrição	Exemplos	Operações
Nominal	Os valores de um atributo nominal são apenas nomes diferentes, atributos nominais provem somente informações o suficiente para se distinguir objetos entre si. (=, ≠)	CEP, número do ID de empregados, cor dos olhos, sexo (masculino, feminino)	Moda, entropia, contingência correlacional.
Ordinal	Os valores de um atributo ordinal provê informações o suficiente para ordenar objetos (<, >)	Dureza de minerais (bom, melhor, ótimo), notas, número de ruas.	Média, percentil, correlação de classificação.
Intervalo	para atributos de intervalo, as diferenças entre valores são significativas, existe uma unidade de medida (+, -)	Datas, temperatura em celsius ou fahrenheit	Média, desvio padrão, correlação de pearson, teste t e F.
Ratio/Razão	para variáveis de razão/proporção, ambas as diferenças e proporções são significativas (*, /)	Temperatura em Kelvins, quantidades monetárias, contagem, idade, massa, comprimento.	Média geométrica, media harmónica, variação percentual.

Compreensão de Dados

- Os dados podem ser resumidos e avaliados
  - Variáveis podem ser numéricas, strings (texto) e datas.
    - Garanta que os dados estão corretamente tipificados: CEPs são números, mas não são numéricos.
- Variáveis continuas
  - Os valores podem, em princípio, variar de infinito negativo a infinito positivo
  - Inteiros ou reais
    - Idade, renda, lucro/perda de lucro, valor de uma fatura, contagem de visitas, dias desde a última visita, são todos variáveis continuas.
  - Alguns valores da variável podem ser restritos de alguma forma (idade).
- Variáveis categóricas
  - Número limitado de valores -> Rotular a variável ao invés de medir
    - Estado, cor de uma bola, raça de um cão.
- Variáveis binárias ou flag
  - Variáveis categóricas contendo apenas dois valores.
    - Repostas a uma questão (Sim ou Não), variáveis fictícias.

Compreensão de Dados

- Núcleo do estágio de compreensão de dados
  - Estatística resumida/básica
  - Visualização de Dados
  - Ganho de insights com os dados
    - Os dados são bons?
    - Estão limpos?
    - Representam aquilo que supostamente deve ser medido?
    - Estão distribuídos como esperado?
    - Vai ser útil para construir modelos?

Estatística resumida/básica

- Maneira mais simples de conseguir insights
  - Estatística básica
    - Média, desvio padrão, skewness, kurtosis.

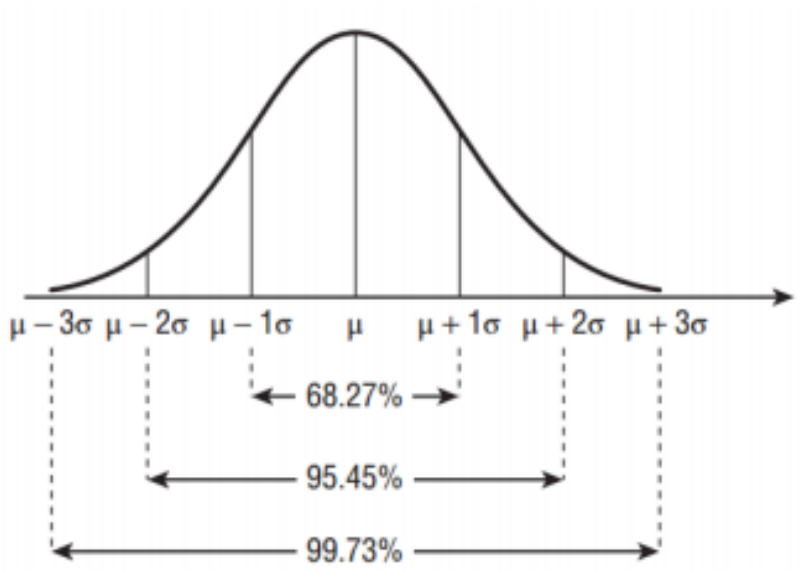
- Média (μ)
  - soma de todos os valores da variável dividido pela contagem de quantos valores a variável possui.
  - meio da distribuição ou um valor típico.
    - Verdade, quando uma variável tem uma distribuição normal ou uniforme.

- Desvio Padrão
  - O desvio padrão é uma medida que expressa o grau de dispersão de um conjunto de dados.
    - Um desvio padrão maior significa que a distribuição de valores para a variável tem um intervalo maior.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

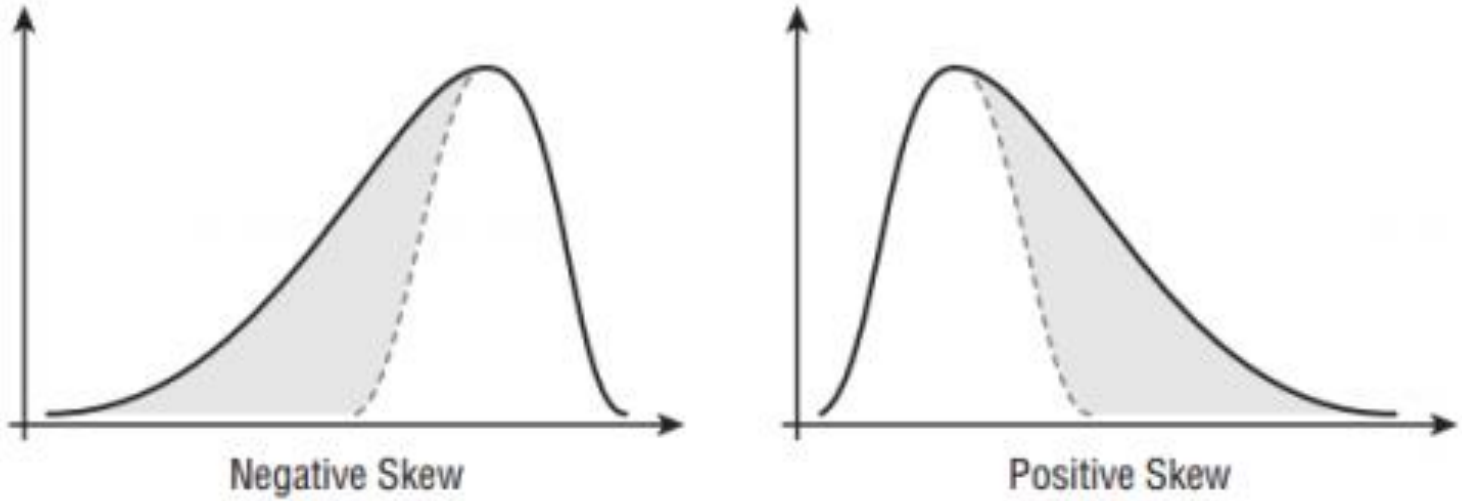
Estatística resumida/básica

- Distribuição Normal
  - Muitos algoritmos assumem distribuições normais.
  - Distribuição Normal
    - A distribuição é simétrica.
    - O valor médio é o valor mais provável de ocorrer no distribuição.
    - A média, mediana e modo são todos do mesmo valor.
    - Aproximadamente 68% dos dados estarão entre a média e ± 1 desvio padrão da média.
    - Aproximadamente 95% dos dados estarão entre a média e ± 2 desvios-padrão da média.



Estatística resumida/básica

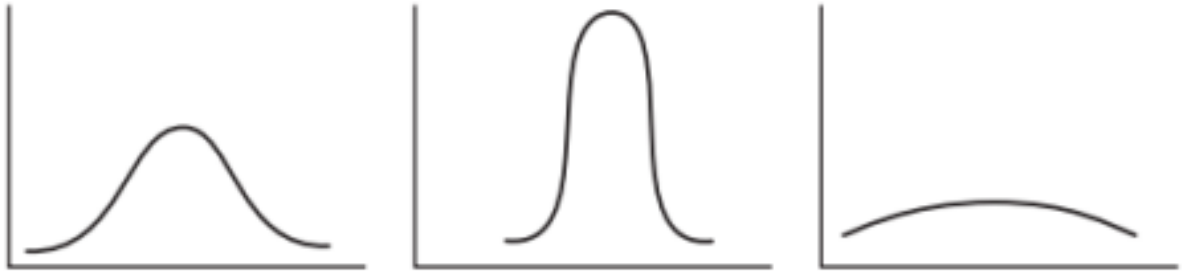
- Skewness
  - Mede o quão balanceada a distribuição é.
  - A Distribuição Normal tem um skewness de valor 0.
  - Inclinação Positiva
    - A distribuição tem uma calda para a direita do corpo principal da distribuição.
  - Inclinação Negativa.
    - A distribuição tem uma calda para a esquerda do corpo principal da distribuição.



Estatística resumida/básica

- Kurtosis
  - Mede quanto mais fino ou mais gordo a distribuição é comparada as distribuições normais.
  - A Distribuição Normal tem uma kutosis de valor 3.
  - Leptokurtic (kurtosis de valor <3)
  - Platykurtic (kurtosis de valor >3)

$$kurtosis = \frac{\sum (X_i - \mu)^4}{N\sigma^4} - 3$$



Visualização de Dados

- **Visualização de dados** - o processo de exibição de dados (geralmente em grandes quantidades) de maneira significativa para fornecer insights que suportarão melhores decisões.
  - A visualização de dados melhora a tomada de decisões, fornece aos gerentes melhores capacidades de análise que reduzem a dependência de profissionais de TI e aprimora a colaboração e o compartilhamento de informações.
  - Visualização de dados é o processo de conversão de dados em figuras simples de interpretáveis.

"transformation from numbers to insight requires two stages."

Jacques Bertin in Semiology of Graphics"

Visualização de Dados

- Benefícios
  - Veja diferentes perspetivas dos dados.
  - Faz com que a interpretação de grandes quantidades de dados seja possível.
  - Encontre exceções nos dados.
  - Permite análise de padrões visuais.
  - Permitir que os analistas passem por dados e se orientem visualmente para os padrões nos dados.

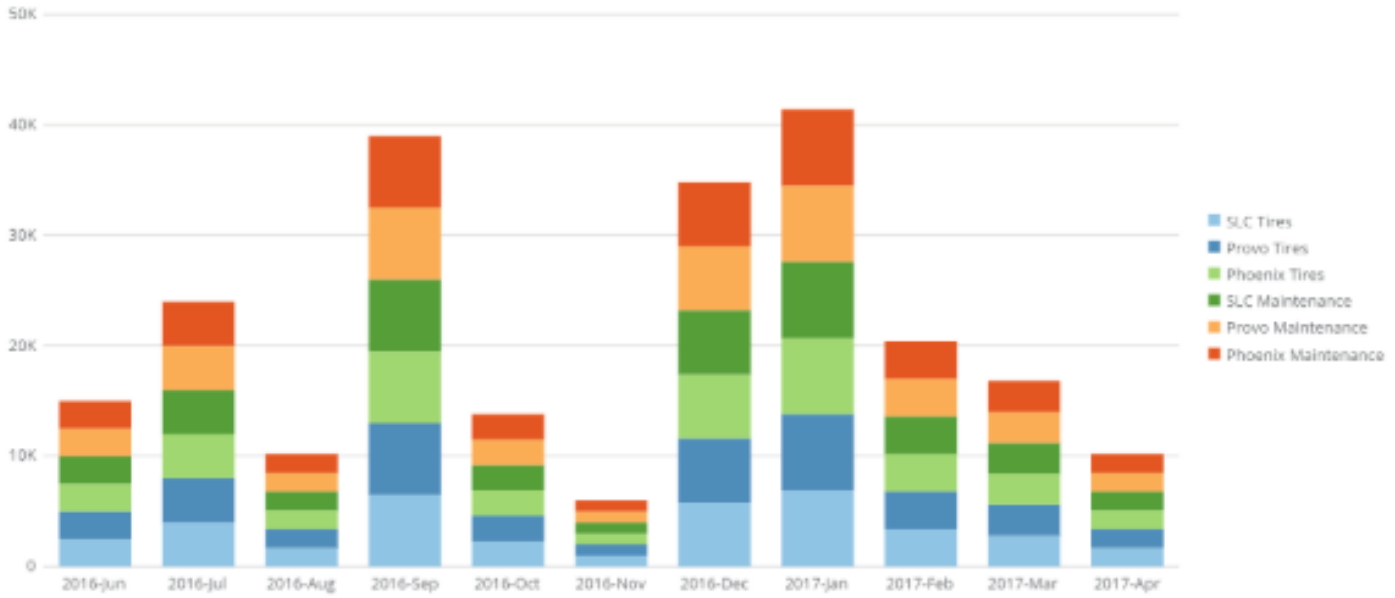
Análise de dados Tabular vs. Visual

- Dados tabulares podem ser usados para determinar exatamente quantos unidades de um determinado produto foram vendidos em um particular mês, ou para comparar um mês para outro.

	A	B	C	D	E	F
1	Month	Product A	Product B	Product C	Product D	Product E
2	January	7792	5554	3105	3168	10350
3	February	7268	3024	3228	3751	8965
4	March	7049	5543	2147	3319	6827
5	April	7560	5232	2636	4057	8544
6	May	8233	5450	2726	3837	7535
7	June	8629	3943	2705	4664	9070
8	July	8702	5991	2891	5418	8389
9	August	9215	3920	2782	4085	7367
10	September	8986	4753	2524	5575	5377
11	October	8654	4746	3258	5333	7645
12	November	8315	3566	2144	4924	8173
13	December	7978	5670	3071	6563	6088

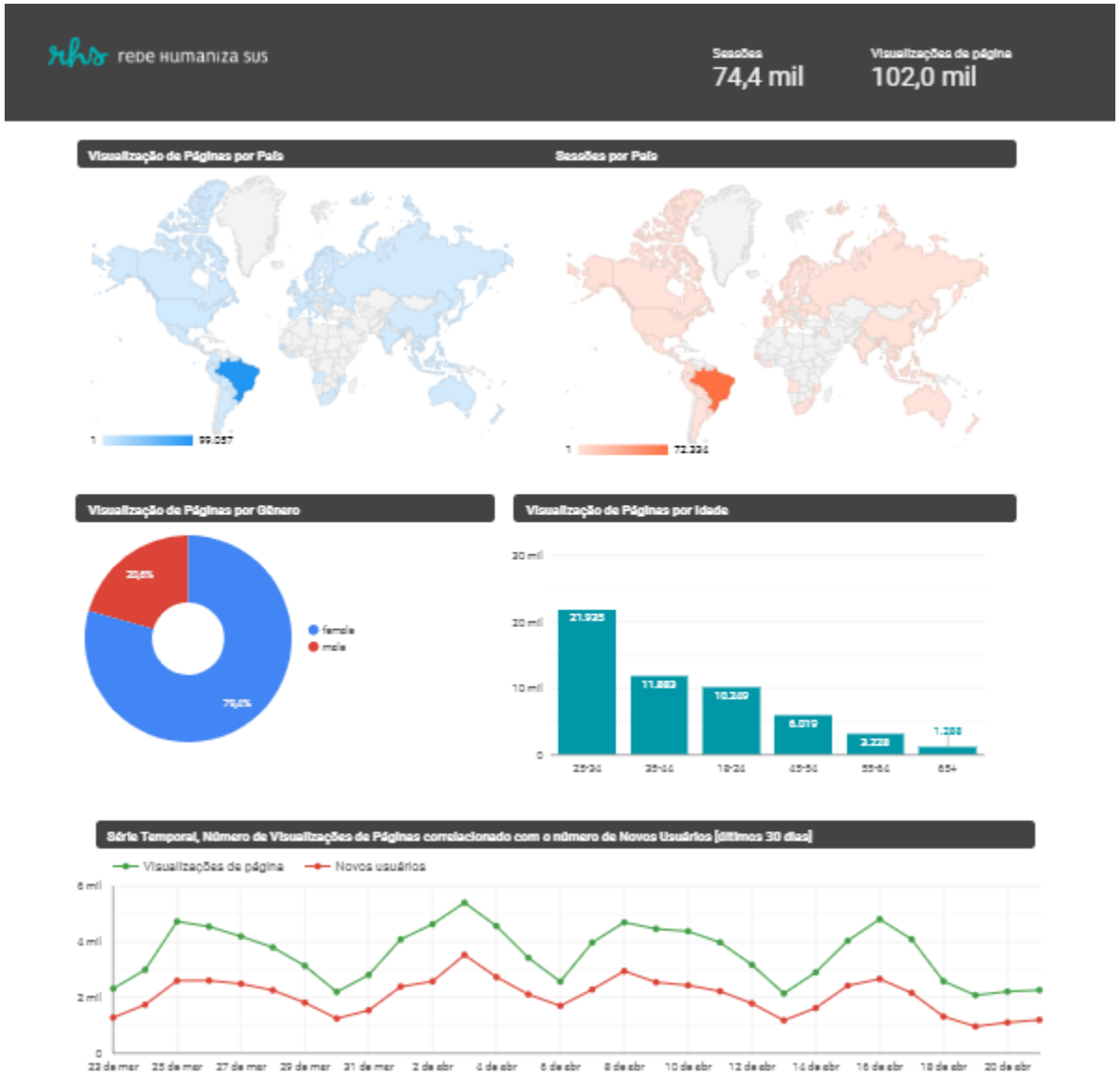
Análise de dados Tabular vs. Visual

- Um gráfico visual fornece os meios para
  - comparar facilmente as vendas globais de diferentes produtos
  - identificar tendências
  - outros padrões

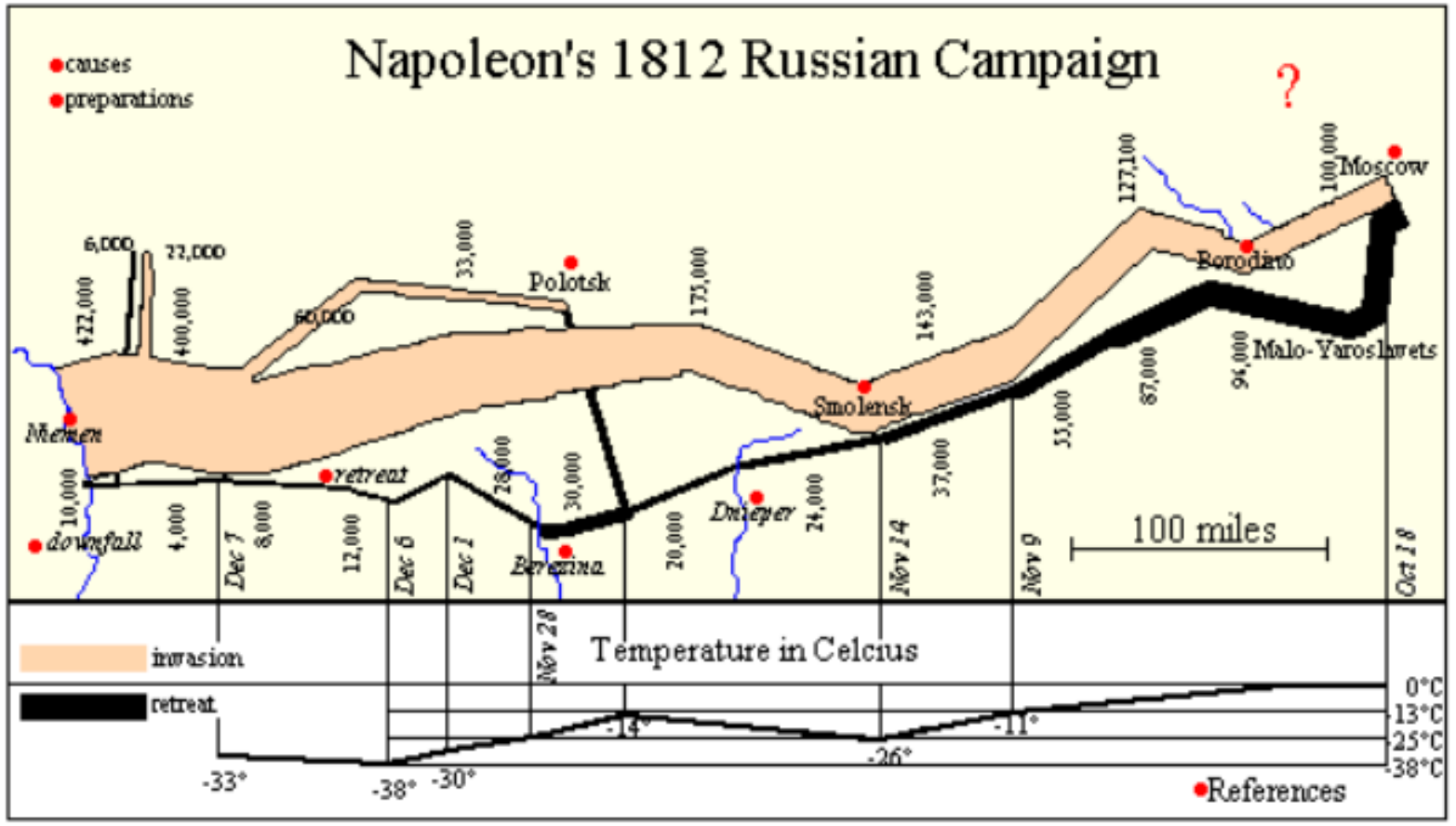


Análise de dados Tabular vs. Visual

- Dashboards

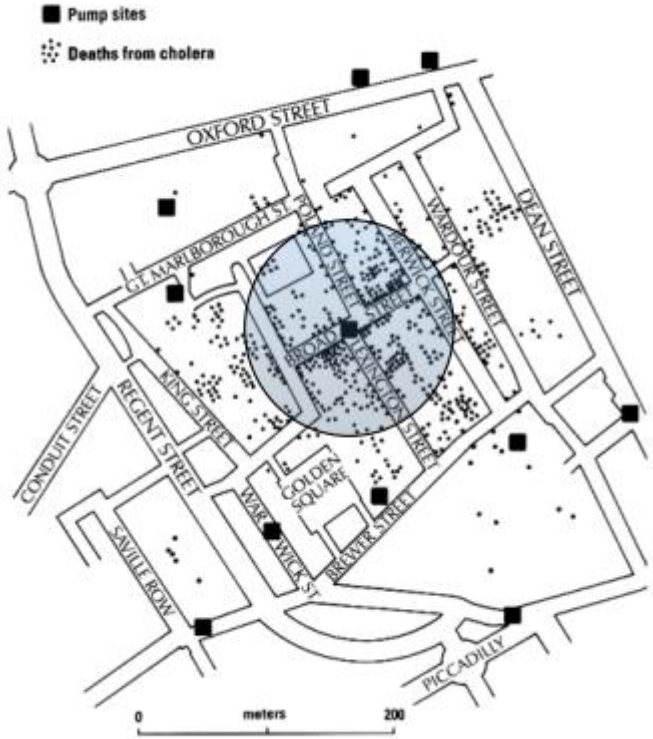


Exemplos



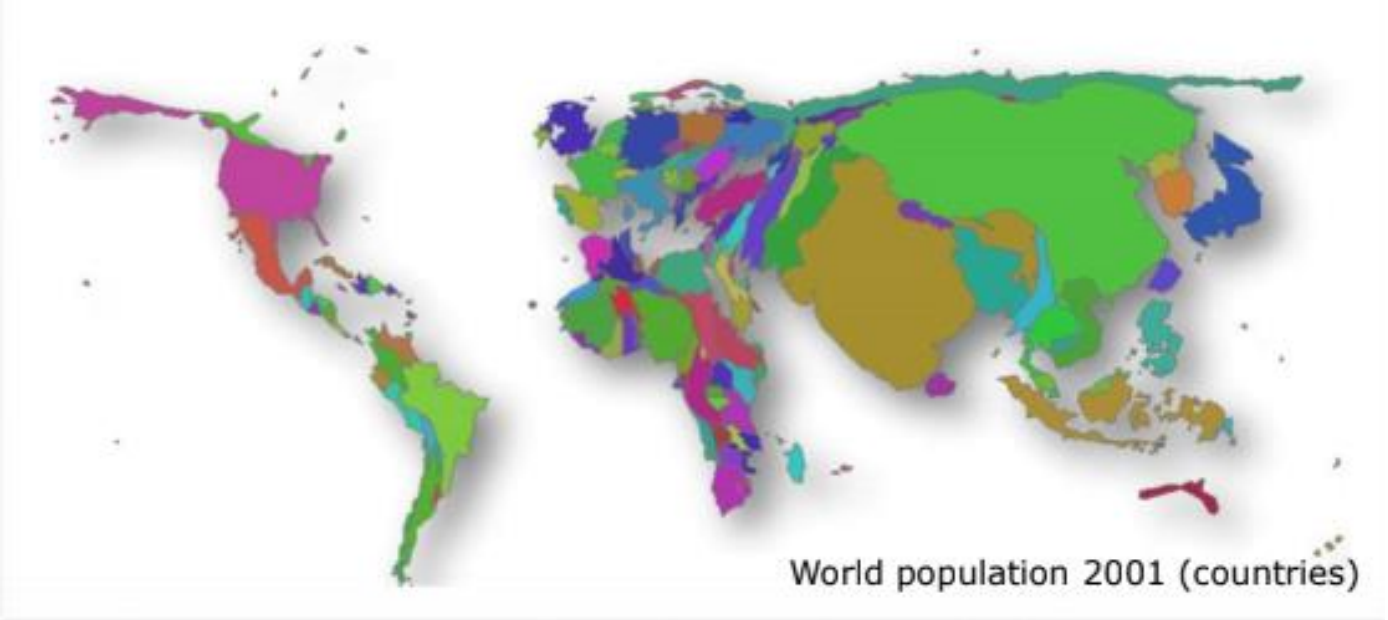
http://www.math.yorku.ca/SCS/Gallery/minard/march-animated.gif

Exemplos



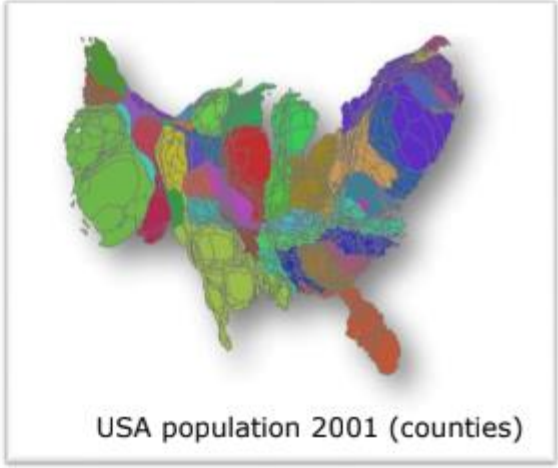
Mapa da Cólera Reino Unido.

Exemplos



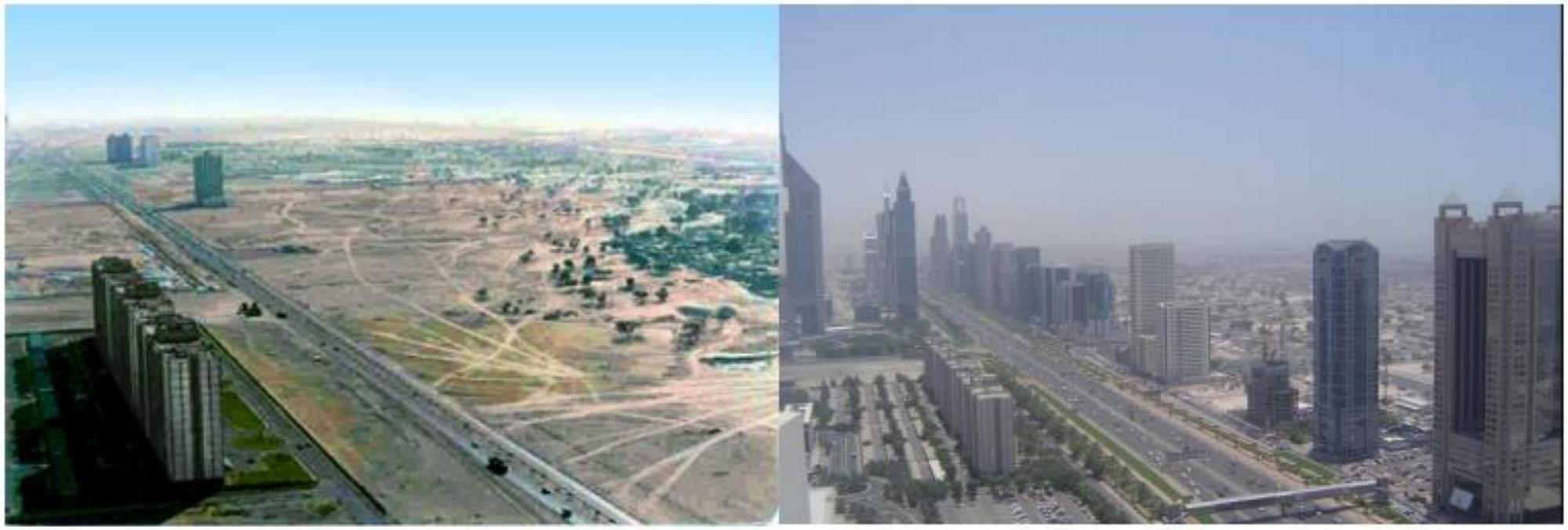
World population 2001 (countries)

USA population 2001 (state)



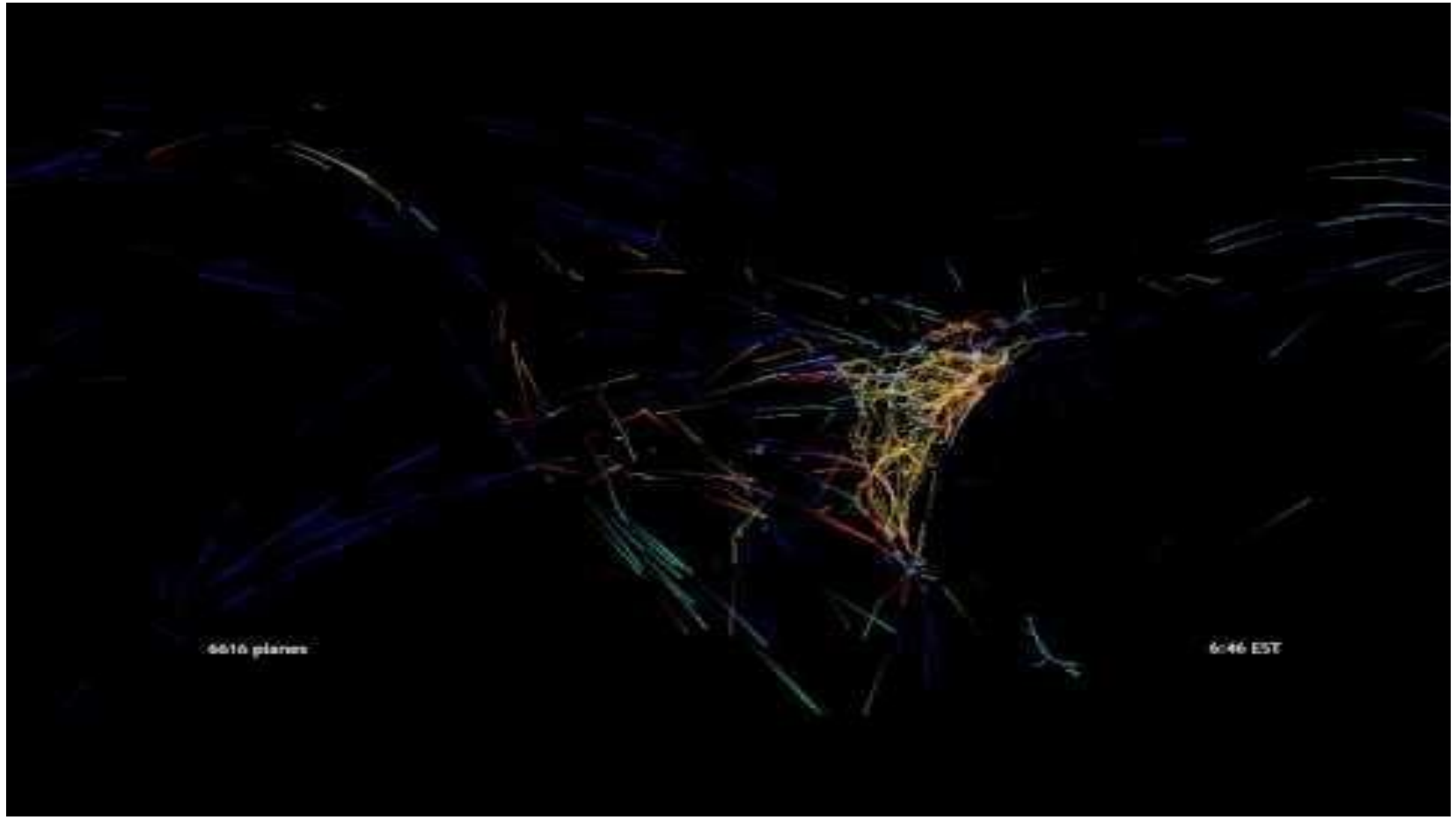
USA population 2001 (counties)

Exemplos



Dubai 1991 vs 2005

Exemplos



https://www.youtube.com/watch?v=ystkKXzt9Wk

Análise Exploratória de Dados

- **Teste de hipóteses versus análise exploratórios de dados.**
  - Analista pode ter hipótese “a priori” para testar
    - o aumento da estrutura de taxas levou à redução da participação de mercado?
- Testando hipóteses
  - Teste a hipótese de que quota de mercado diminuiu.

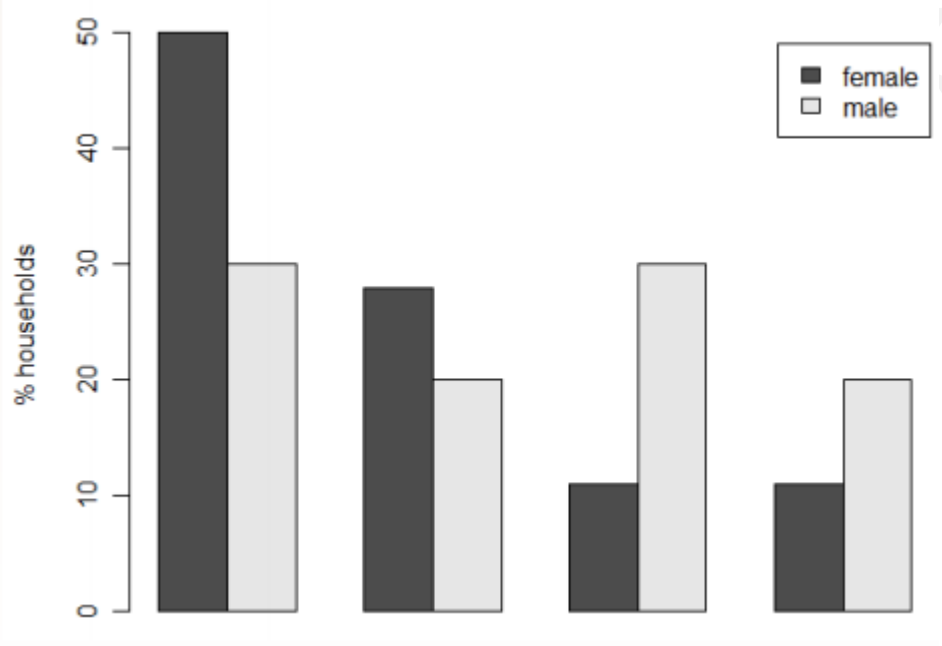
- Contudo
  - Com grandes e desconhecidos bancos de dados, analistas às vezes não tem uma hipótese
  - Análise exploratória de dados (AED)
    - examinar as inter-relações entre os atributos;
    - identificar subconjuntos interessantes das observações;
    - desenvolver uma ideia de possíveis associações entre os preditores, bem como entre os preditores e o alvo.

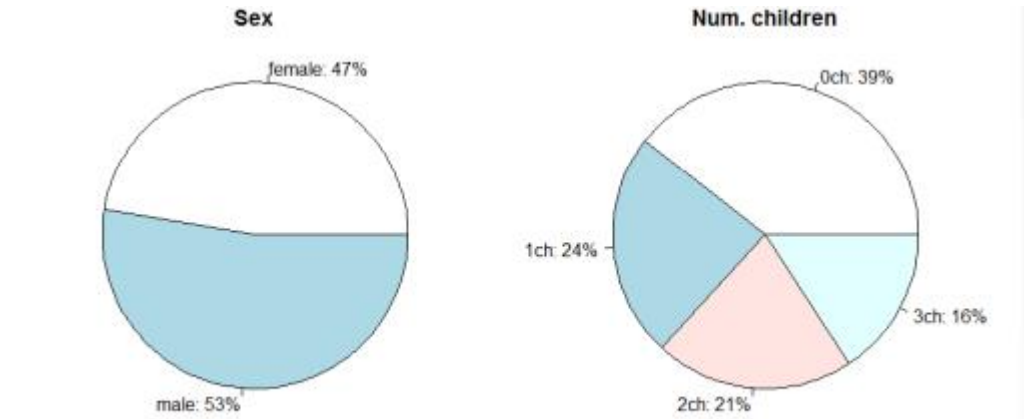
Análise Exploratória de Dados

A AED vai além do uso descritivo da estatística, procura olhar de forma mais profunda os dados, sem resumir muito a quantidade de informações.

Portanto, a qualidade na representação gráfica deve ser pautada na **clareza, simplicidade e autoexplicação**. As técnicas gráficas desempenham um papel fundamental na AED.

variável qualitativa*	variável quantitativa
tabela de frequências gráfico de barras diagrama circular (pizza)	medidas de posição: média, mediana, moda medidas de dispersão: variância, desvio-padrão, amplitude, coeficiente de variação tabela de frequências histograma boxplot gráfico de linha ou sequência polígono de frequências





http://www.each.usp.br/lauretto/SIN5008_2011/aula01/aula1

Análise de Dados Exploratória

- Churn data set
  - Churn data set. Blake, C.L. & Merz, C.J. UCI Repository of machine learning databases. Irvine, CA: University of California, Department of Information and Computer Science 1998

Variable	type	Obs
State	Categorical	for the 50 states and the District of Columbia
Account Length	Integer-valued	how long account has been active
Area code	Categorical	
Phone Number	Essentially a surrogate for customer ID	
International Plan	Dichotomous categorical	yes or no
Voice Mail Plan	Dichotomous categorical	yes or no
Number of Voice Mail Messages	Integer-valued	
Total Day Minutes	Continuous	minutes customer used service during the day
Total Day Calls	Integer-valued	
Total Day Charge	Continuous	perhaps based on above two variables
Total Eve Minutes	Continuous	minutes customer used service during the evening
Total Eve Calls	Integer-valued	
Total Eve Charge	Continuous	perhaps based on above two variables
Total Night Minutes	Continuous	minutes customer used service during the night
Total Night Calls	Integer-valued	
Total Night Charge	Continuous	perhaps based on above two variables
Total International Minutes	Continuous	minutes customer make international calls
Total International Calls	Integer-valued	
Total International Charge	Continuous	perhaps based on above two variables
Number of Calls to Customer Service	Integer-valued.	
Churn	Target.	customer has left the company (True or False)

Referências

- Churn data set
  - Churn data set. Blake, C.L. & Merz, C.J. UCI Repository of machine learning databases. Irvine, CA: University of California, Department of Information and Computer Science 1998
- http://www.math.yorku.ca/SCS/Gallery/minard/march-animated.gif
- http://www.aaronkoblin.com/work/flightpatterns/
- https://mineracaodedados.files.wordpress.com/2012/04/the-crisp-dm-model-the-new-blueprint-for-data-mining-shearer-colin.pdf
- http://www.each.usp.br/lauretto/SIN5008_2011/aula01/aula1
- https://paulovasconcellos.com.br/crisp-dm-semma-e-kdd-conhe%C3%A7a-as-melhores-t%C3%A9cnicas-para-explora%C3%A7%C3%A3o-de-dados-560d294547d2
- https://metodosdigitaisufg.wordpress.com/2018/02/24/analise-de-dados-utilizando-python/

Obrigado

Real Python