



Descoberta do Conhecimento





Descoberta do Conhecimento

Cleilton Lima Rocha

Universidade 7 de Setembro
Fortaleza - CE, Brasil



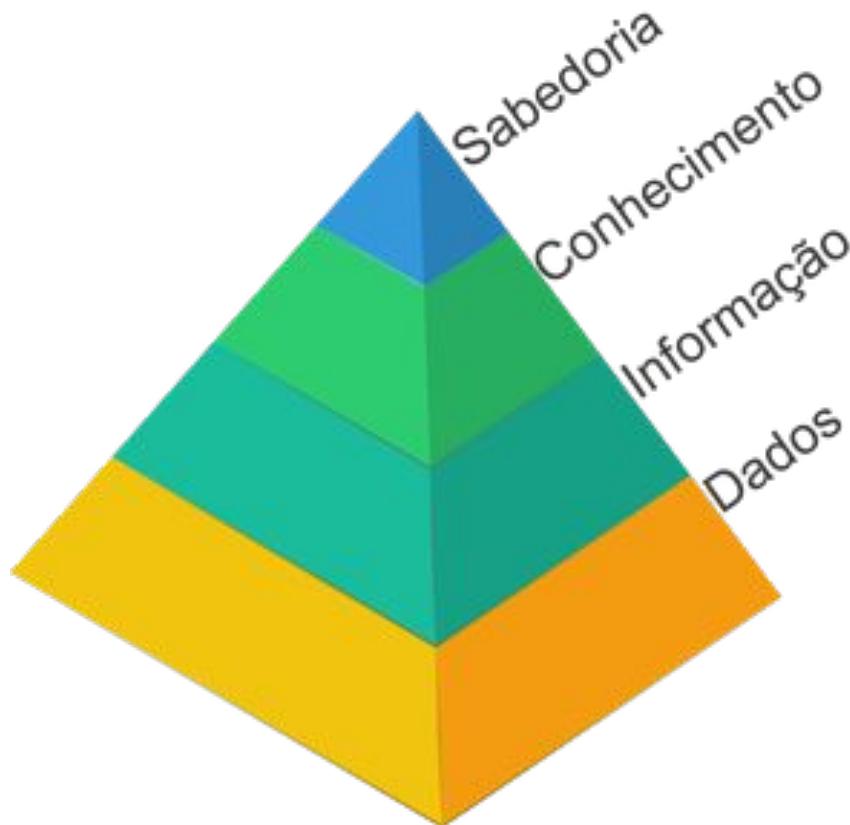


Agenda

- ◊ Introdução ao Processo de Descoberta de Conhecimento e Data Science
 - ◊ Engenharia de Atributos:
 - Pré-processamento de dados
 - Seleção de Features ...
 - ◊ Aprendizagem supervisionada
 - Classificação
 - Regressão
 - ◊ Aprendizagem não supervisionada
 - ◊ Análise do bias variance threshold
 - ◊ Projeto prático aplicado à Data Science.
- 



Processo de Descoberta de Conhecimento





“O KDD pode ser visto como o processo de descoberta de padrões e tendências por análise de grandes conjuntos de dados, tendo como principal etapa o processo de mineração, consistindo na execução prática de análise e de algoritmos específicos que, sob limitações de eficiência computacionais aceitáveis, produz uma relação particular de padrões a partir de dados FAYYAD et al (1996).”



*“Informação é o resultado do processamento de dados num formato que tem significado para o usuário respectivo e que tem **valor real ou potencial** nas **decisões** presentes ou prospectivas DAVIS (1974).”*

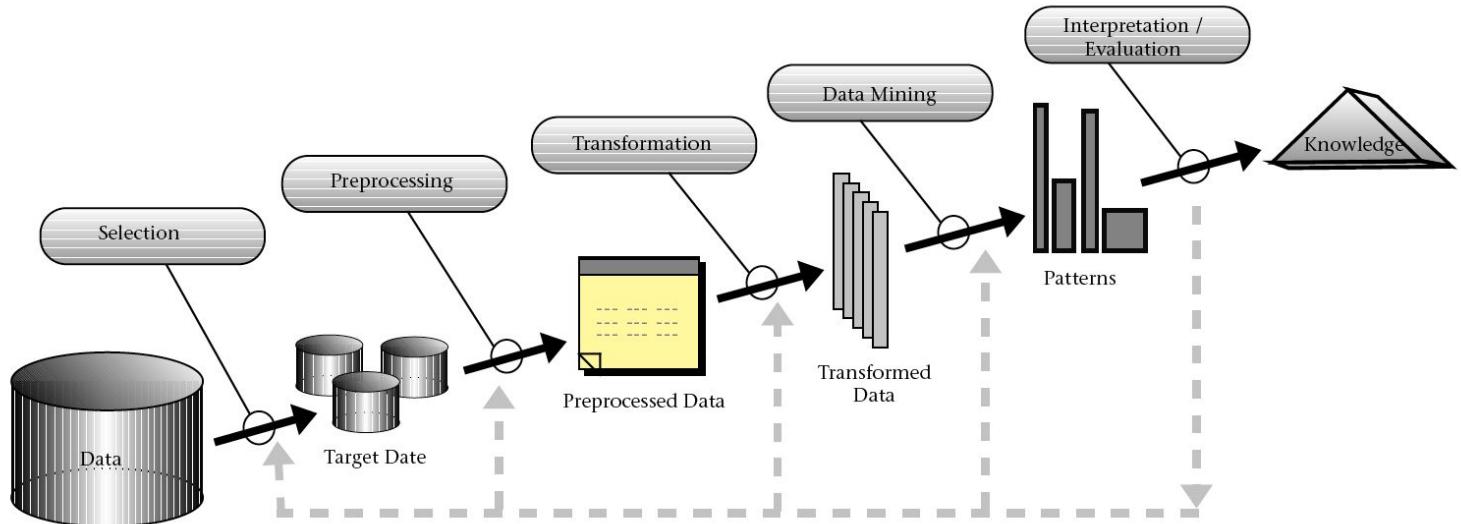


“O conhecimento necessário para se decidir e/ou avaliar torna-se disponível por meio de informações SANCHES (1997).”

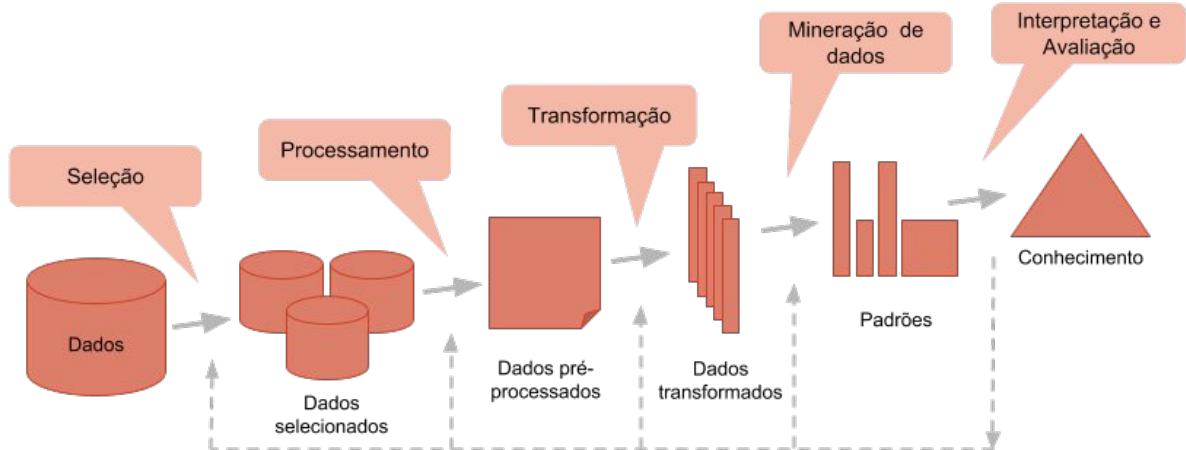


*“Segundo DAVENPORT e PRUSAK (1998), a gestão do conhecimento (GC) pode ser vista como **uma série de ações gerenciais constantes** e sistemáticas que facilitam os processos de criação, registro e compartilhamento do conhecimento nas organizações.”*

Fases do KDD



Fases do KDD





Data Mining e seus métodos

- ◊ Aprendizagem supervisionada
- ◊ Aprendizagem não supervisionada
- ◊ Modelos de regras de associação
- ◊ Modelos de relacionamento entre variáveis



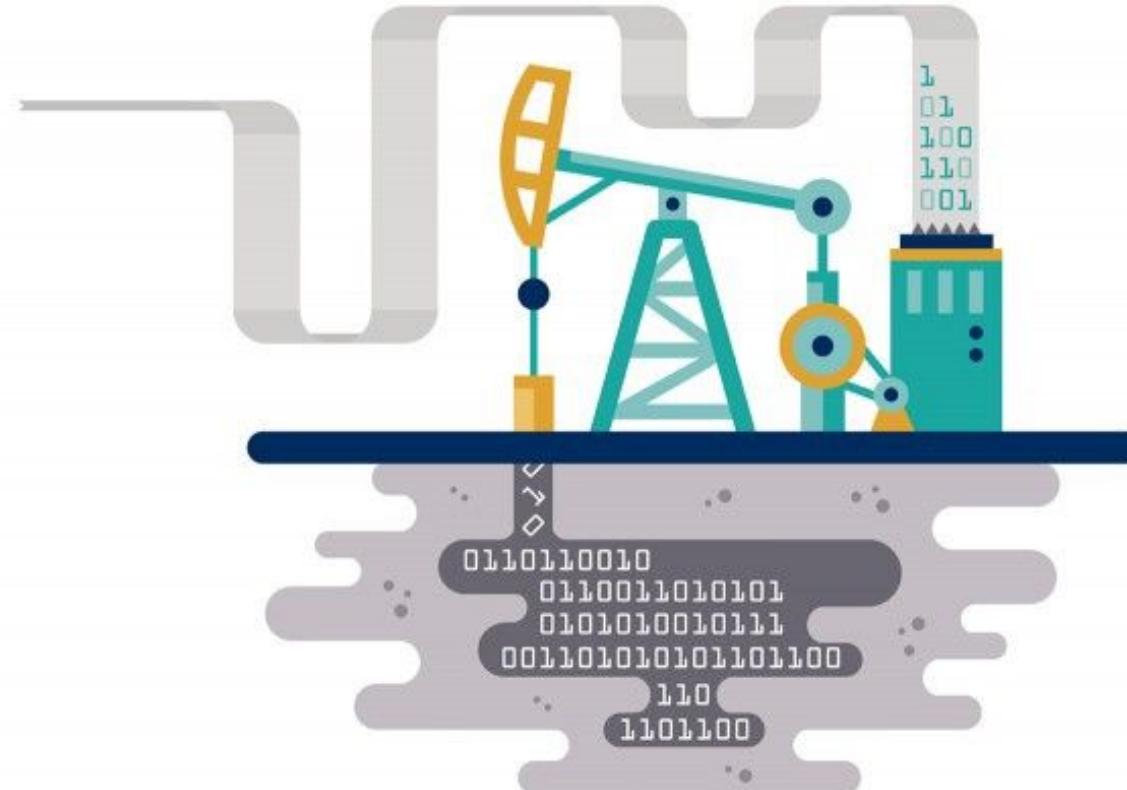
ATD

Apoio à tomada de decisão



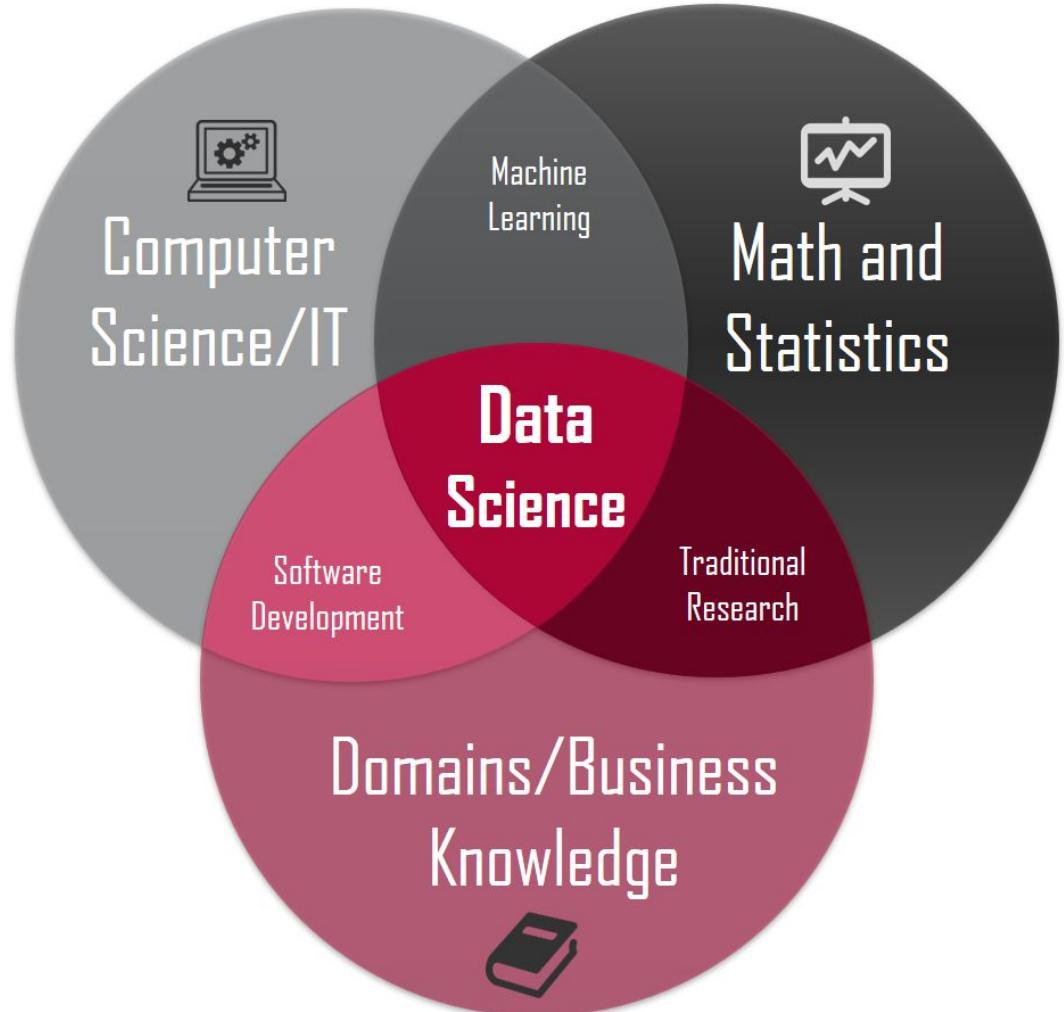


Riqueza dos Dados



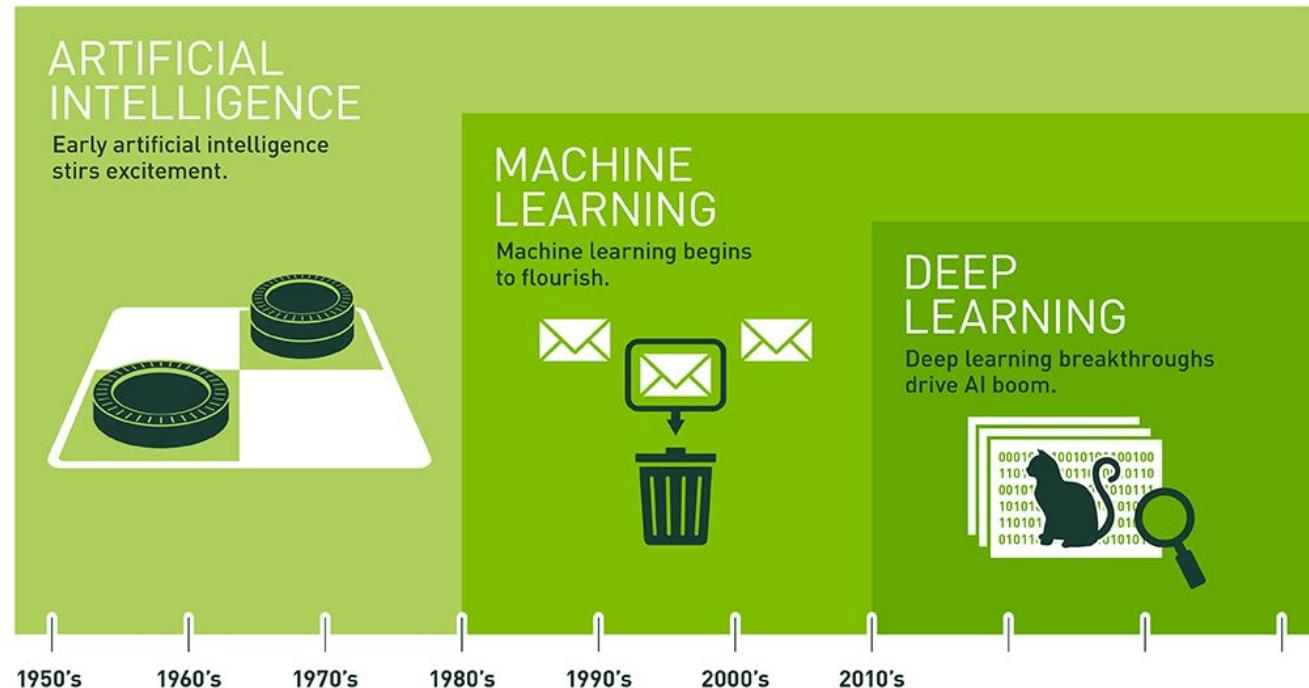


Data Science – uma ciência
interdisciplinar





Machine Learning Overview



Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.



Data Science Overview

data **quantitative** **statistical** **inference** **models** **statistics**

research coefficient regression learning generalized linear bayesian probability modeling maximum research coefficient regression learning generalized linear bayesian probability modeling maximum

analytics expectation likelihood trend management spatial visualization methods predictive normal parameter time function causal equation simulation workshops consulting covariate duration variance distribution graphical standard



Exemplos



Recommendation Systems



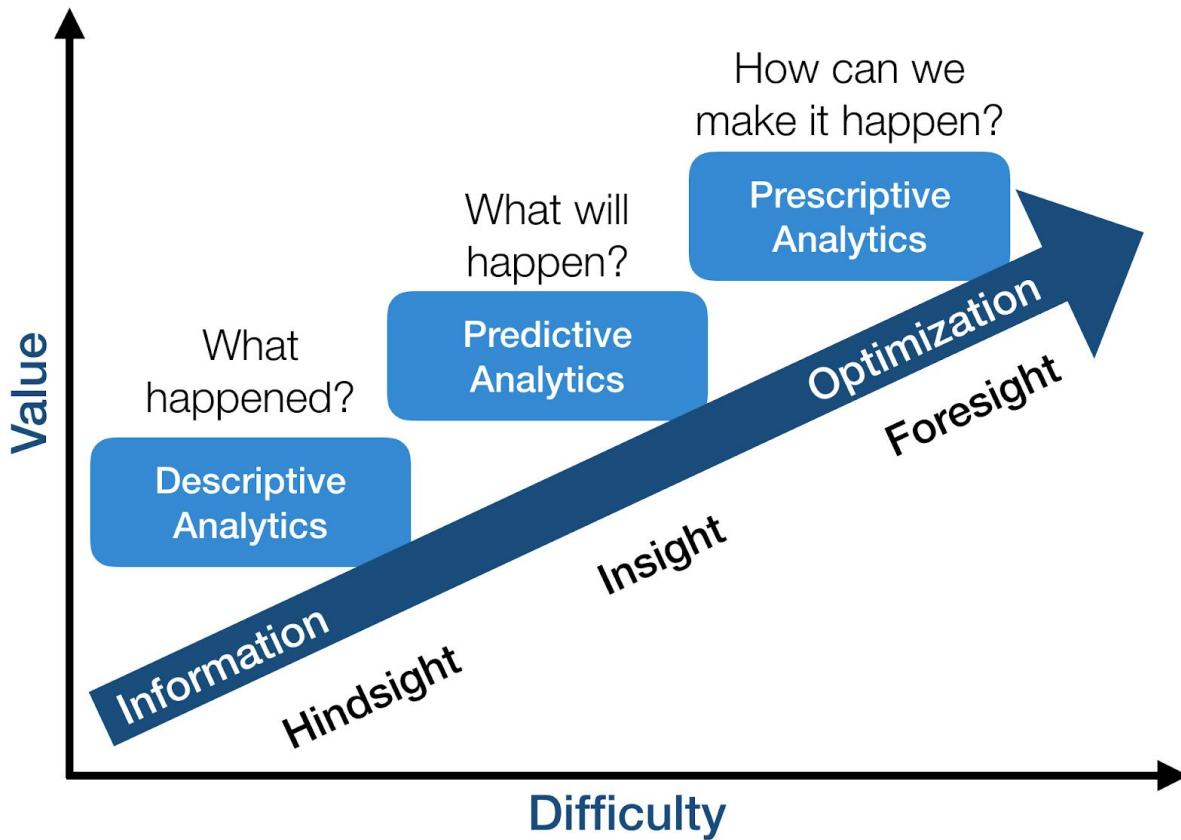
Inventory planning



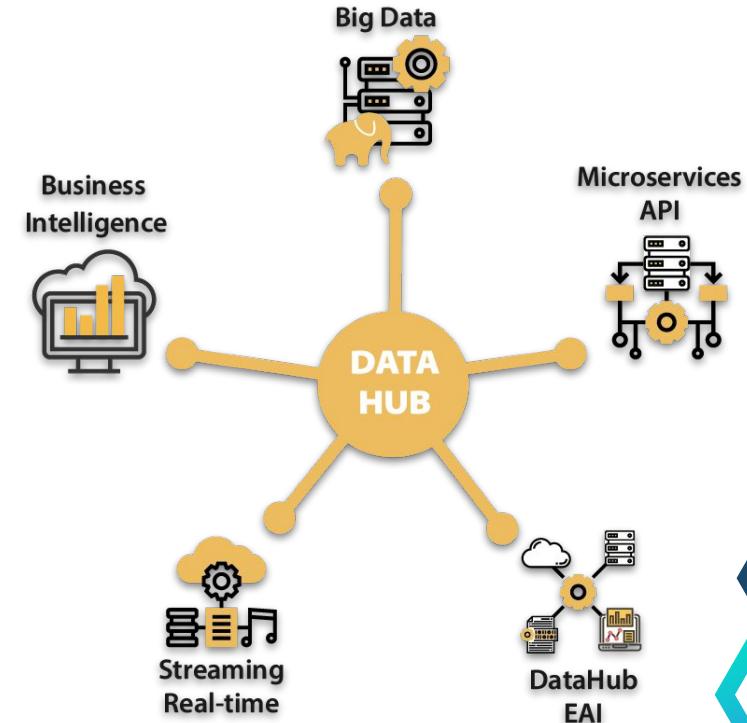
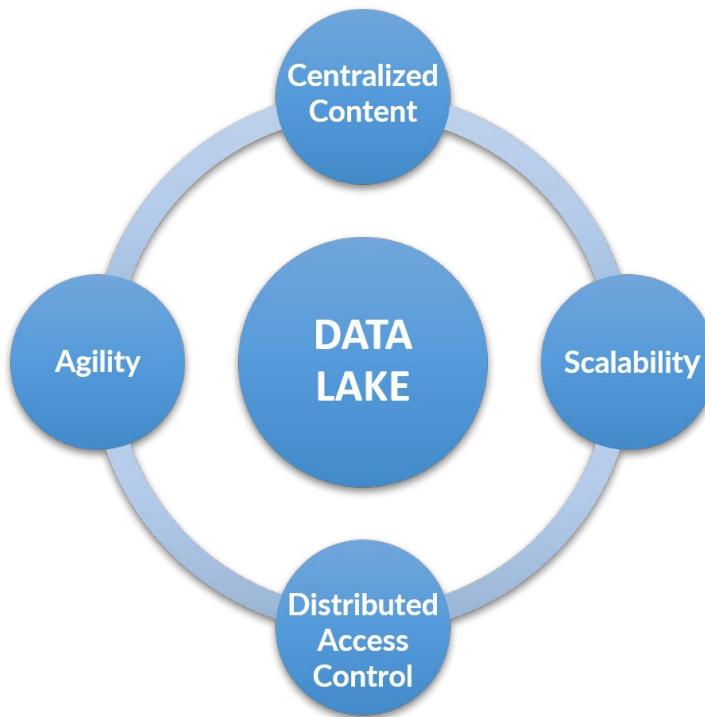
Dynamic
pricing



Cadeia de Valor



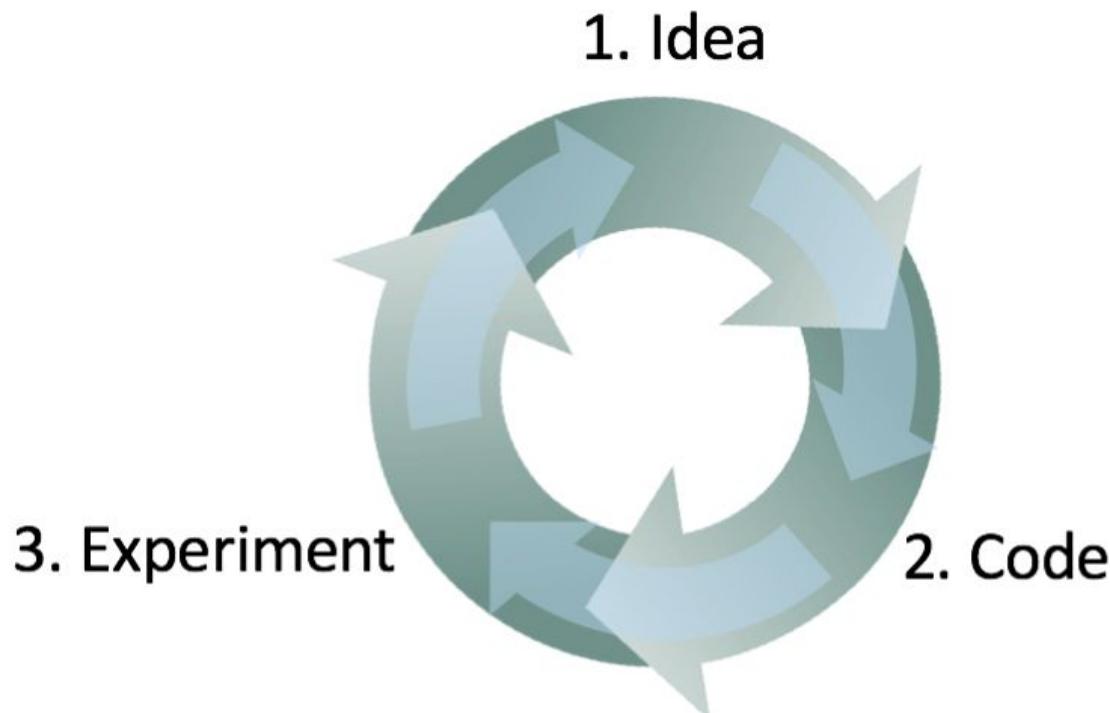
Integração de dados massiva





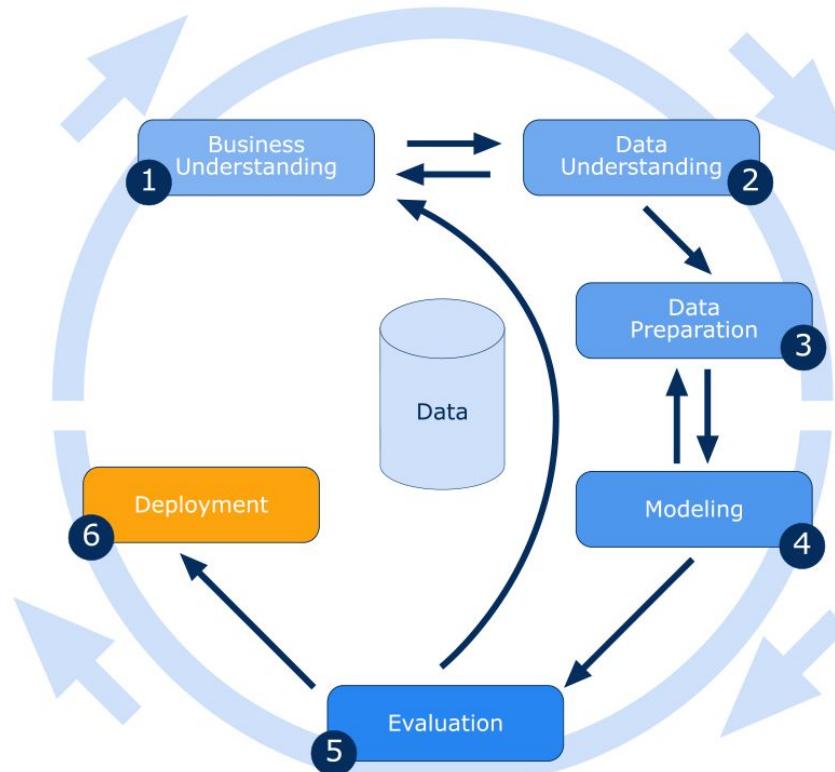


Metodología



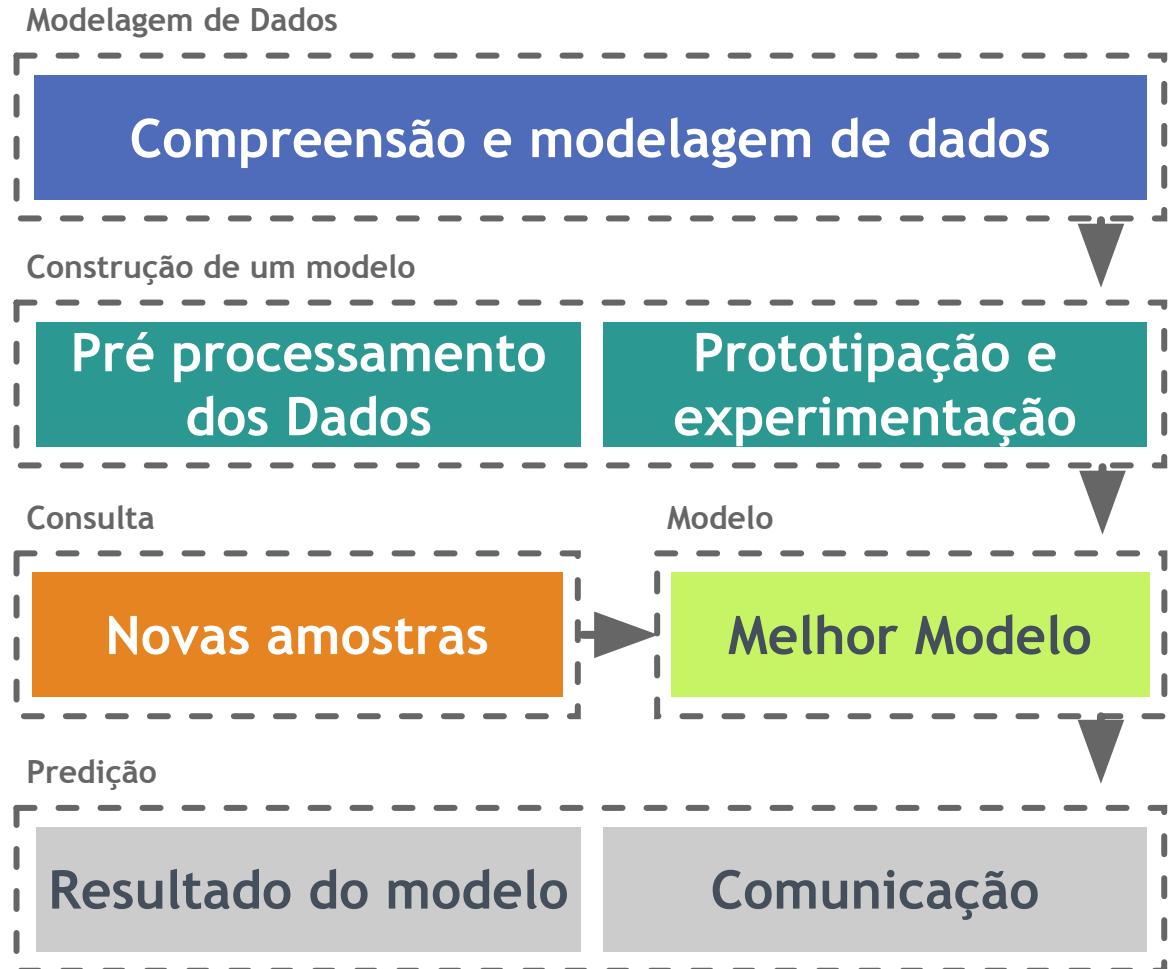


Metodología





Metodologia





Exploração de Dados



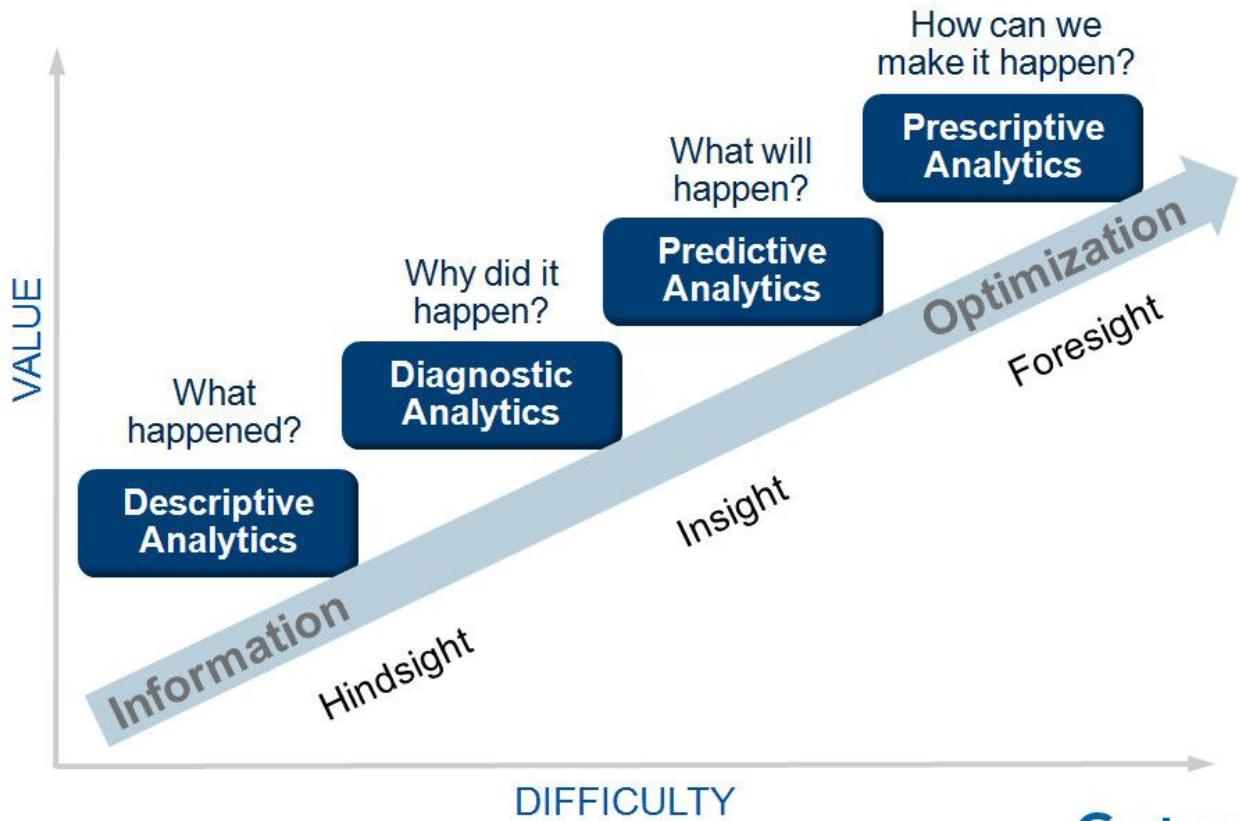
Medidas estatísticas





Exploração de dados

Analytic Value Escalator





Medidas estatísticas

Utilizaremos as medidas estatísticas para resumir e analisar uma conjunto de dados.

Para tal, temos:

- **Medidas de posição ou tendência central:** focaremos nas medidas de tendência central, onde se verifica a tendência dos dados em termos de valores centrais.
- **Medidas de dispersão:** análise de como os dados se distribuem e variam conforme os valores centrais.

As medidas podem ser aplicadas para analisar tanto a população quanto os dados de uma amostra. Porém, trabalhar com **população** é mais difícil, dada as **restrições de tempo, custo e processamento**, por isso na maioria dos casos trabalharemos com **amostras**.



Medidas estatísticas

Para que possamos entender o significado das fórmulas, precisamos entender o conceito de uma **série de dados**.

Portanto, sendo **X** uma variável e x_1, x_2, \dots, x_n os possíveis valores de **X** em um dado instante, temos que **X** representa uma **série de dados**.

Por exemplo:

Série X: 1, 3, 5 e 7; Temos $x_1 = 1$, $x_2 = 3$, $x_3 = 5$ e $x_4 = 7$ com $n = 4$;





Medidas de tendência central

Iremos estudar as seguintes medidas:

- Média
 - Moda
 - Mediana
-
- **Média aritmética simples:** valor para onde se concentram os dados da distribuição;

$$\bar{X} = \frac{\sum xi}{n}$$

- **Média aritmética ponderada:** ponderada cada valor sugerindo que cada um tem seu peso diferente no valor da média final;

$$\bar{X} = \frac{\sum xi \times pi}{\sum pi}$$



Medidas de tendência central

- **Média geométrica:** A média geométrica sugere interpretações geométricas, como o nome sugere. Mas também é muito utilizada para encontrar a média em valores que crescem proporcionalmente. Portanto, utilizada em aplicações financeiras e na geometria;

$$M_g = \sqrt[n]{x_1 \times x_2 \times x_3 \dots x_n}$$

- **Média harmônica:** Utilizada para cálculos de média que compreendem grandezas inversamente proporcionais.

$$M_h = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} \dots \frac{1}{x_n}}$$



Medidas de tendência central

- **Mediana:** medida de tendência quando desejamos determinar o valor que separa a série de dados em duas partes iguais;

Para encontrar o valor central, temos duas situações:

Se n for ímpar, será o valor de posição:

$$M_{ed} = \frac{n+1}{2}$$

Se n for par, será a média dos dois valores centrais:

$$M_{ed} = \bar{X}\left(\frac{n}{2}, \frac{n+1}{2}\right)$$

Observe que nesse caso o valor da mediana pode não representar valores da série.

Medidas de tendência central

- **Moda:** representa o valor mais frequente da série de dados;
 - Considere a série: 1,2,1,5,5,1. Qual a **moda**?

Valor	1	2	5
Frequência	3	1	2





Medidas de dispersão

Há cenários em que a **média** não é suficiente para caracterizar uma série de dados.

Exemplo: Imagine que há um vendedor de produtos. No primeiro cenário, ele almeja vender nos 3 primeiros meses do ano a seguinte quantidade de produtos: 10, 15, 20 produtos para o primeiro, segundo e terceiro mês, respectivamente. Agora suponha o segundo cenário: 5, 10, 30. Qual destes representa o cenário mais estável de ganhos?

Se calcularmos a **média**, o resultado será **igual a 15**. No entanto, o primeiro cenário representa o mais estável em torno deste valor central.

Iremos discutir como caracterizar tais cenários com as **medidas de dispersão**.



“A **média** é o valor que melhor representa uma série de valores, mas ela, por si só, não pode destacar o grau de homogeneidade ou heterogeneidade existente entre os valores que compõem o conjunto.” [1]

Variância

- **Variância:** quantifica a dispersão dos dados. Portanto, permite identificar a homogeneidade ou heterogeneidade de uma série de dados.
 - Para séries **homogêneas**: menor variância;
 - Para séries **heterogêneas**: maior variância;

$$s^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$$



Desvio Padrão

- **Desvio padrão:** assim como a variância, o desvio padrão quantifica a dispersão de dados em relação à média. No entanto, elimina a amplificação dos desvios segundo a variância. Logo, teremos a percepção do quanto distante os valores da série estão distantes da média.
- $S = \sqrt{S^2}$

$$S = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$



Coeficiente de variação de Pearson

E para calcular a dispersão de duas séries distintas?

	Média	Desvio Padrão
Estatura	175 cm	5,0 cm
Peso	68 kg	2,0 kg

Perceba que há dois tipos de séries com diferentes unidades de medida. Como afirmar qual conjunto de dados tem maior variabilidade?

Coeficiente de variação de Pearson

$$CV = \frac{s}{\bar{x}} \times 100$$

- Para o exemplo anterior:

$$\text{CVP estatura} = (5/175) \times 100 = 2,86\%$$

$$\text{CVP peso} = (2/68) \times 100 = 2,94$$

Percebemos que o peso possui maior CVP, apresentando então maior variabilidade dos dados.

Esse resultado demonstra uma característica interessante. Vamos analisar a tabela novamente.

Coeficiente de variação de Pearson

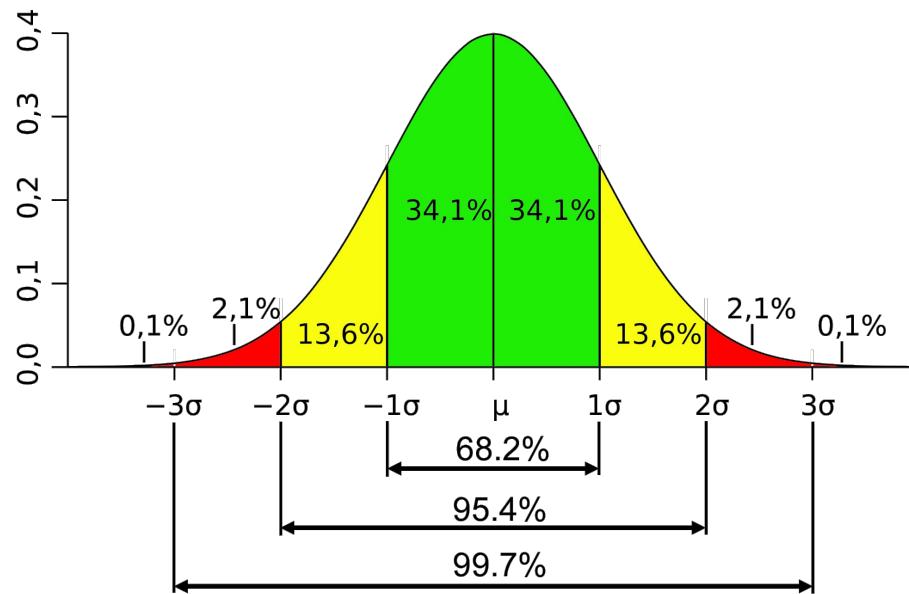
	Média	Desvio Padrão
Estatura	175 cm	5,0 cm
Peso	68 kg	2,0 kg

Perceba que mesmo o **peso** apresentando **menor desvio padrão**, possui **maior variabilidade**. Isso se dá pela relação do desvio padrão em relação à média, onde uma variação de 2 unidades no peso representa uma variação maior que 5 unidades na altura.

Portanto, a CVP é uma ótima medida para cenários de comparação como esse de séries.

Distribuição Normal

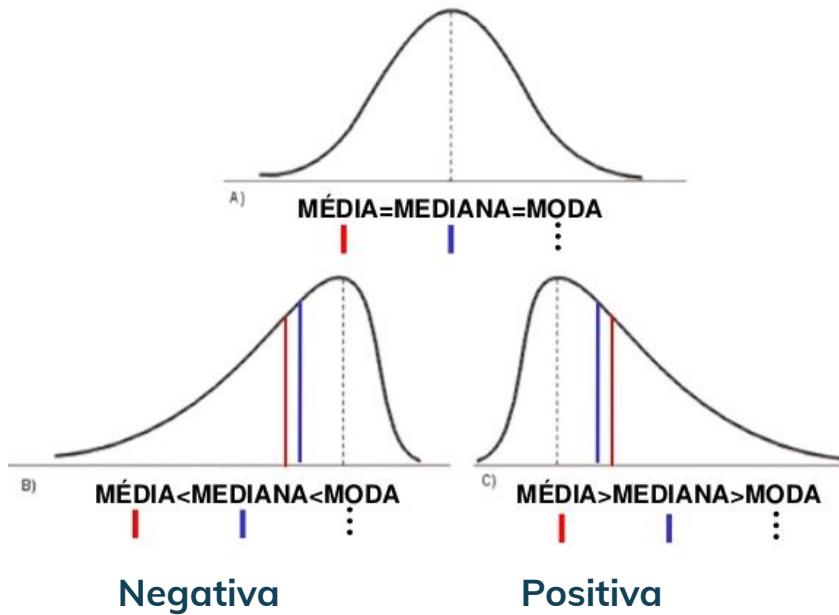
Com a curva normal definida, temos informações importantes sobre a distribuição dos nossos dados:



Intervalo	Proporção
$\mu \pm 1\sigma$	68,2%
$\mu \pm 2\sigma$	95,4%
$\mu \pm 3\sigma$	99,7%

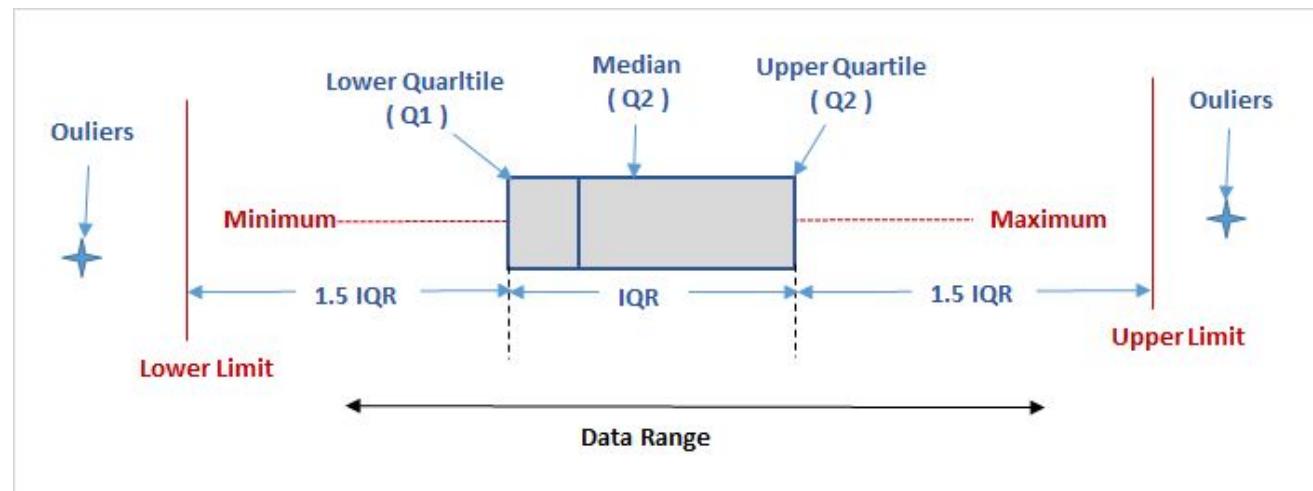
Distribuição

Simetria



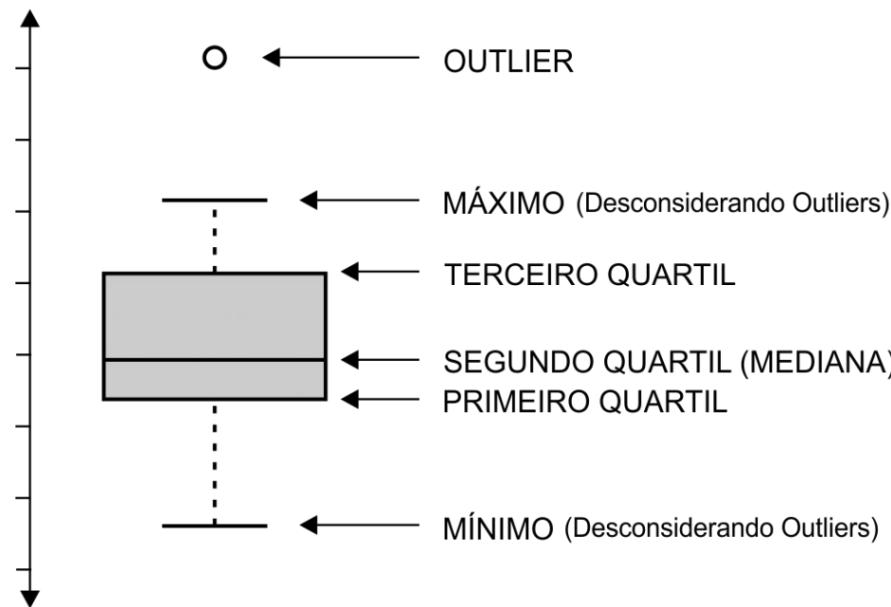


Box Plot



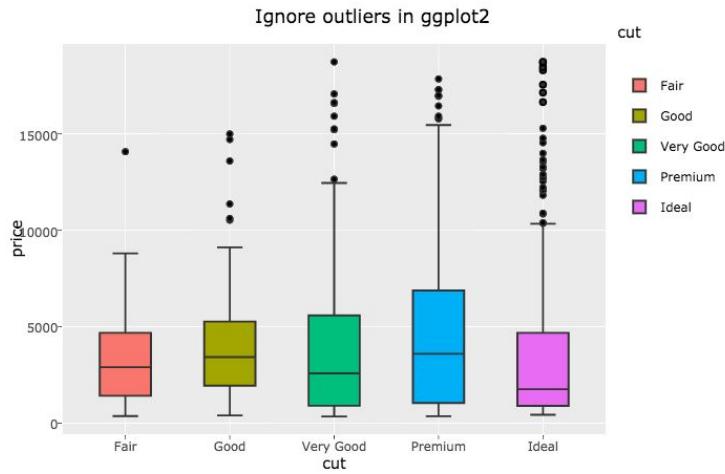
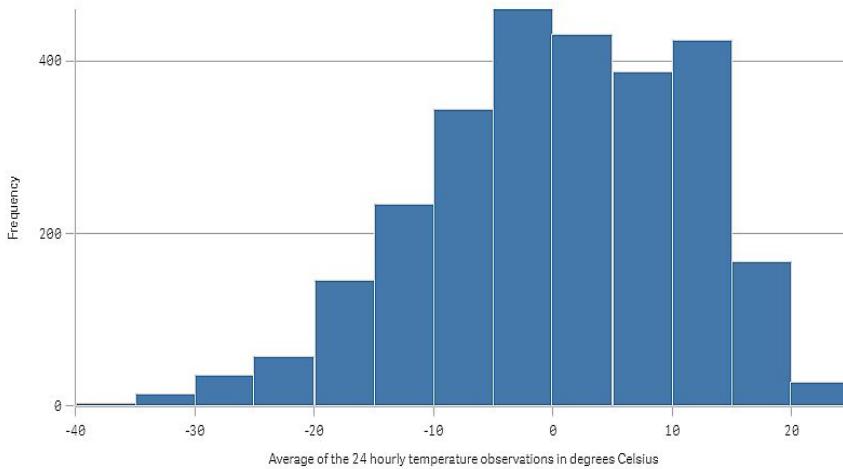


Box Plot



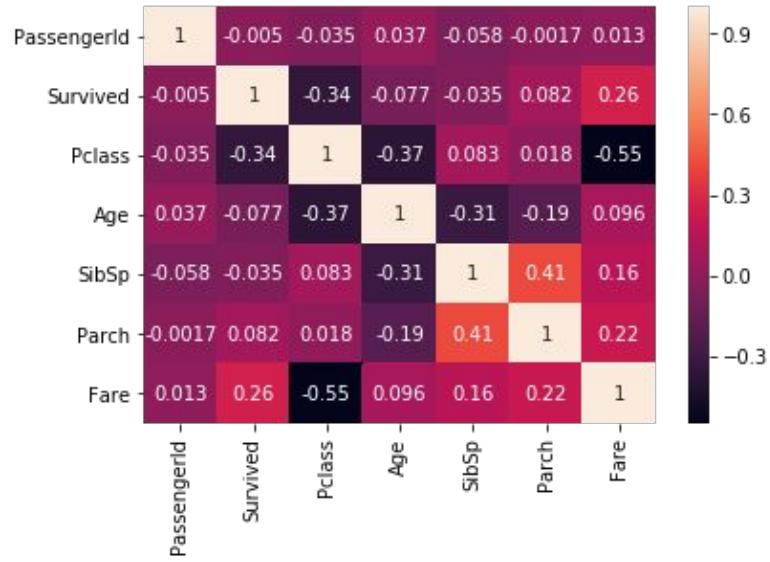
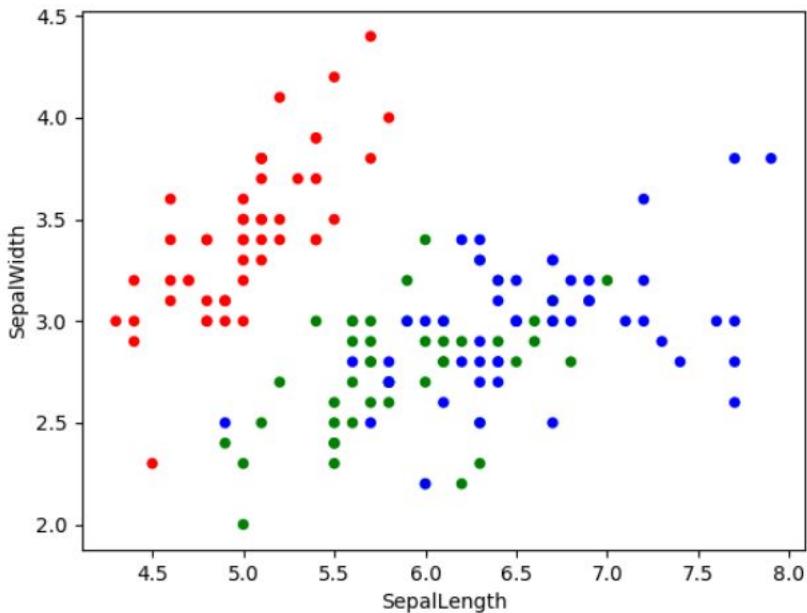


Exploração do dado





Exploração do dado





http://dontpad.com/kdd_uni7

Hands-On

