## Introdução à Análise Exploratória e os Ambientes R

Nesta aula você aprenderá:

* O que é Análise Exploratória de Dados e sua Importância
* Onde se insere e Análise Exploratória de Dados no contexto geral de Ciência de Dados
* O que é a linguagem R e como empregar seus ambientes programação

# Introdução

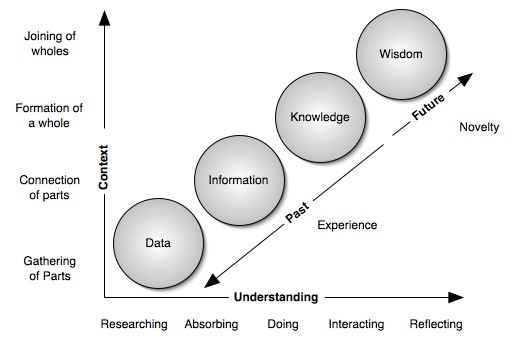
A ampla disponibilidade de dados obtida com a transformação digital (pense aqui particularmente nas empresas e negócios) leva a um grande interesse pela busca de métodos para extrair informações úteis e conhecimento de dados que podem então ser empregados para a tomada de decisões, por pessoas – como decidir por uma estratégia de vendas ou de produção para os próximos meses – ou automáticas – como a recomendação de produtos em um site de e-commerce ou o bloqueio de uma transação de cartão de crédito suspeita de fraude. O potencial de empregarmos dados para a tomada de decisões é de grande importância para os negócios e os dados um importante ativo das empresas e, por esse motivo, você talvez você já tenha encontrado frases do tipo *os dados são o novo petróleo das empresas*.

Em quaisquer casos, o que está por traz disso é um pensamento analítico sobre os dados (*Data Thinking*), e que se encontra no centro do que chamamos Ciência de Dados. Uma parte importante desse processo de Análise e Ciência de Dados é a Análise Exploratória de Dados que você irá começar a conhecer aqui.

# Dados, Informação e Conhecimento

Uma diferenciação importante que precisamos fazer é entre os conceitos de dados, informação e conhecimento. Dados são os registros brutos armazenados e podem ser entendidos como a matéria-prima das informações. São o primeiro estágio de um processo de construção de conhecimento. Você pode pensar nos cliques dos usuários em um site Web que são armazenados. Dependendo do site podem ser milhares de registros por hora. A informação corresponde aos dados tratados e organizados que podem ser analisados de maneira a virar conhecimento. Pense, por exemplo, em organizarmos os dados anteriores em páginas mais e menos acessadas, nos acessos por hora etc. Por fim, o conhecimento corresponde ao terceiro nível, quando temos a informação transformada em experiência. Enquanto a informação é como um bem comum, que pode ser compartilhado igualmente entre várias pessoas, o conhecimento é a informação transformada pelo indivíduo em uma assimilação crítica da informação. A informação das páginas mais acessadas pode levar o indivíduo a entender as características de páginas mais atraem os usuários. Mas aqui, cada um, empresa ou indivíduo, dá uma direção diferente para o uso da informação.

Para dar mais um exemplo, pense em uma pesquisa de marketing de uma empresa de produtos de beleza. Os registros coletados pelos pesquisadores são os dados. Os dados da pesquisa são então tabulados e organizados e ficam disponíveis para consulta. Eles fornecem agora informação. Isso ser tornará conhecimento na medida que essas informações forem analisadas e utilizadas para a definição de ações efetivas de marketing. Por exemplo, a pesquisa pode indicar que a marca tem pouca penetração nos jovens por não estar associada a aspectos de diversidade ou ecologia e, assim, encaminhar uma mudança nas campanhas de seus produtos.



A figura acima mostra como os conceitos de dados, informação e conhecimento estão relacionados hierarquicamente e apresenta ainda um nível hierárquico ainda maior, o de *sabedoria*. Esse nível, entretanto, tem mais interesse para estudos específicos de teoria do conhecimento e, para o uso prático de Análise de Dados podemos nos limitar aos conceitos de dados, informação e conhecimento.

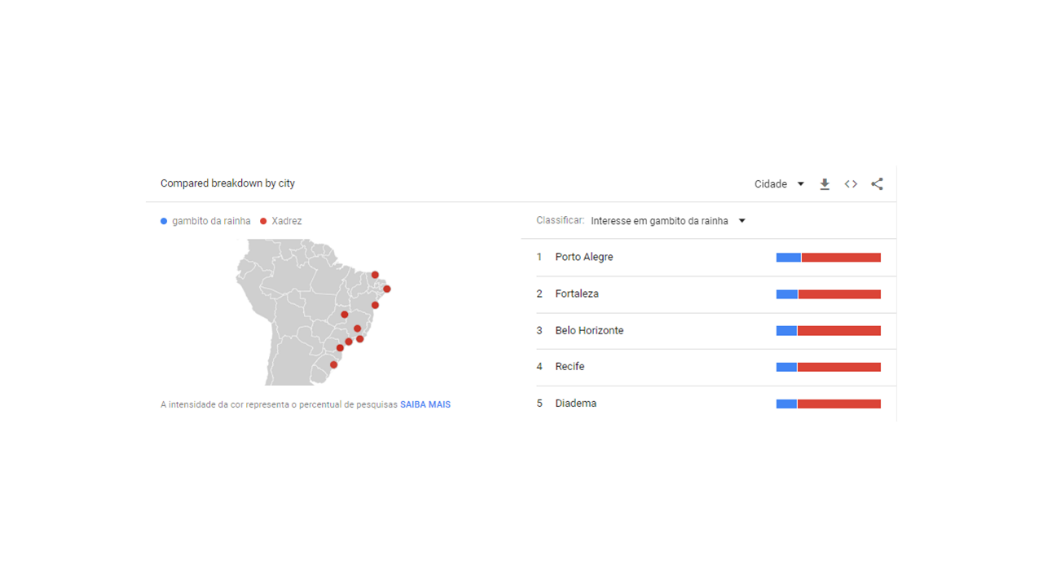
# Um exemplo prático: Google Trends

O [Google Trends](https://trends.google.com.br/) (<https://trends.google.com.br/>) é uma ferramenta que permite analisar dados sobre a pesquisa de termos na Web pelo navegador da Google. Nela os dados coletados (os termos pesquisados em todo o mundo através do Google) já encontram-se na forma de informação, já estando organizados por data, local, tipos e termos de busca. Você pode então usar essa informação para, por exemplo, pesquisar sobre as buscas dos termos *Gambito da Rainha*, a famosa série televisiva, e jogos de *Xadrez* na internet.



Como você pode notar podemos observar um forte aumento nas buscas por jogos de Xadrez a partir do início da famosa série televisiva. Você poderia então empregar esse conhecimento para decidir sobre presentar uma pessoa com um tabuleiro ou um livro de xadrez com grande probabilidade de sucesso, ou uma empresa poderia decidir lançar uma campanha sobre seus jogos de xadrez na internet ou aumentar sua edição de livros de regras de xadrez. Você acaba de fazer sua primeira Análise Exploratória de Dados!

Você ainda pode refinar a sua exploração e verificar o crescimento pela busca de jogos online de xadrez ou as cidades onde as buscas foram mais representativas (abaixo).



Esse tipo de pesquisa com o Google Trends, ou outras ferramentas similares pagas, são de fato usadas para análise de tendência de buscas na internet e permitem inferir sobre o comportamento das pessoas e dos consumidores. A partir daí as empresas podem tomar decisões sobre seus negócios, como o aumento ou a redução da produção de um determinado produto de consumo, a realização de uma campanha de marketing ou mesmo a decisão de criação de um novo produto.

# Clássicos da Análise de Dados

Existem muitos casos clássicos de Análise de Dados que passaram a fazer parte da história e mesmo do folclore da Ciência de Dados dando origem a livros, filmes e uma série de estórias, nem sempre fiéis, mas eles ainda ilustram o potencial e os resultados da aplicação da Análise de Dados. Vale a pena você conhecer alguns desses casos.



* [How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did](https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/#52ff65166668)

(<https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/#52ff65166668>)

Neste case clássico um pai fica indignado ao receber em casa folhetos de promoção de produtos para gestante e bebês julgando que eles expunham sua filha adolescente. Ele não sabia, mas sua filha estava grávida e o Target (uma grande loja de departamentos americana) identificou isso pelas compras que haviam sido feitas com o cartão de crédito do pai.



* [You’ll Never Guess What Surprising Product Walmart Stocks up on Before Big Storms](https://www.southernliving.com/news/walmart-strawberry-pop-tarts-hurricane)

(<https://www.southernliving.com/news/walmart-strawberry-pop-tarts-hurricane>)

Nos períodos de fortes chuvas ou mesmo tornados é comum nos Estados Unidos os cidadãos terem que ficar dias trancados em casa sem poder sair. O Walmart analisando dados dos períodos de chuva de anos anteriores conseguiu verificar que nesses períodos havia uma demanda extraordinária de tortas de maçã (e provavelmente identificou também outros produtos) e pôde, então, se antecipar e provisionar tortas para atender a demanda, potencializando suas vendas e sem causar a falta dos produtos para os clientes.



* [Beer and Diapers (Nappies)](https://bigdatabigworld.wordpress.com/2014/11/25/beer-and-nappies/)

(<https://bigdatabigworld.wordpress.com/2014/11/25/beer-and-nappies/>)

Em mais um case clássico da loja de Departamentos Target, analisando dados do comportamento dos consumidores o Target conseguiu observar uma forte *correlação* de compras de seus clientes: às sextas-feiras, muitos clientes que compravam fraldas também compravam cerveja! Isso levou o Target a colocar esses produtos mais próximos para facilitar a compra dos seus clientes. Mais tarde, descobriu-se que os clientes eram jovens pais que, uma vez que já tinham que ficar sexta à noite cuidando de seus filhos ou em casa com suas esposas, aproveitavam para relaxar assistindo os jogos da temporada na TV.



* [The Story of Moneyball Proves Importance of Both Big Data and Big Ideas](https://datasciencedegree.wisconsin.edu/blog/moneyball-proves-importance-big-data-big-ideas/)

(<https://datasciencedegree.wisconsin.edu/blog/moneyball-proves-importance-big-data-big-ideas/>)

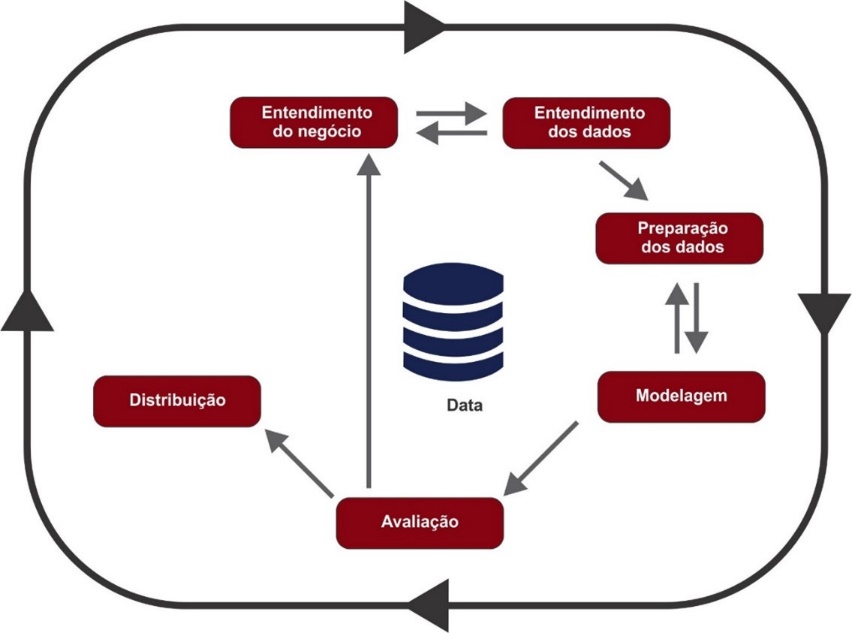
Esse talvez seja o caso mais clássico e por isso tornou-se até mesmo um premiado filme estrelado pelo ator Brad Pitt: Moneyball (O Homem Que Mudou o Jogo). O filme conta a estória de um caso real do impacto do uso de dados para a tomada de decisões. Na temporada de Beisebol de 2001 o Oakland Athletics era um time muito pequeno, passando por dificuldades e sem recursos para contratar grandes estrelas para fazer frente ao campeonato. O diretor do time decide então adotar uma estratégia completamente inovadora. No lugar de tomar decisões com base em conselheiros e *olheiros* de beisebol como todos times faziam à época, ele contrata um estatístico de Yale e passa a decidir sobre a contratação de jogadores e mesmo sobre as posições em que jogariam com base em dados do histórico dos jogadores (quantidade de arremessos com sucesso, pontos por partida etc.). A estratégia conseguiu supreendentemente levar o pequeno time de Oakland até às finais da temporada (algo como levar um time de futebol da terceira divisão à final da libertadores). A partir daí todos os grandes times de Beisebol passaram a empregar estatísticos e analistas de dados para decidir sobre jogadores e estratégias de jogo, e a mesma estratégia foi adotada nos anos seguintes para outros esportes como o Futebol Americano criando uma indústria milionária de análise de dados esportivos.

# O Processo de Ciência de Dados: CRISP DM, Reference Model

Mas esse processo que permite as empresas extrair conhecimento útil dos dados para resolver problemas práticos de seus negócios não é em geral um processo simples. É um processo que pode envolver uma série de técnicas que vão da estatística e ao uso de sofisticados modelos de aprendizado de máquina, passando ainda pelo uso de técnicas eficientes de armazenamento e processamento e muitos outros recursos.

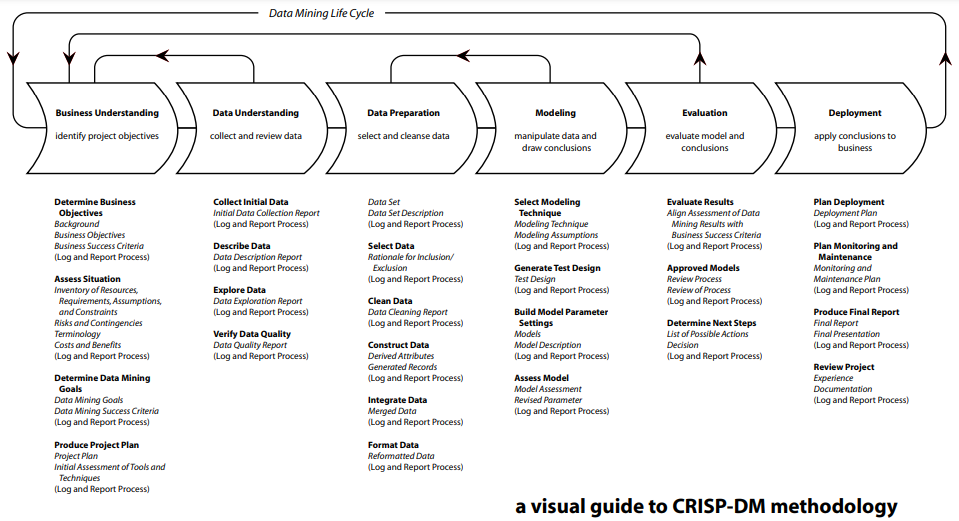
Para lidar com essa complexidade em geral adota-se um processo de fases. O **CRISP-DM, Cross Industry Standard Process for Data Mining** (Processo Padrão Inter-Indústrias para Mineração de Dados é um processo de fases bastante aceito na indústria para representar o ciclo completo de Análise de Dados. Esse processo apresenta algumas variantes ao longo do tempo, como por exemplo a versão do Google em The Process for Data Preparation and Feature Engineering (ver <https://developers.google.com/machine-learning/data-prep/process>), mas de fato elas mudam bastante pouco o modelo CRISP e a essência do processo CRISP-DM é em geral preservada e é empregada na maior parte dos casos.

*Formulação de soluções de mineração de dados e avaliação dos resultados envolve pensar cuidadosamente sobre o contexto em que eles serão usados. Se nosso objetivo é a extração de conhecimento potencialmente útil, como podemos formular o que é útil? Ela depende criticamente da aplicação em questão. (PROVOST, F; FAWCETT, T. Data Science for Business, 2016)*



# As Fases do CRISP-DM

As fases do modelo de referência CRISP-DM encontram-se resumidas no quadro abaixo.



Você pode também querer consultar a documentação original desse modelo de referência.

* [CRISP Visual Guide](http://meusite.mackenzie.br/rogerio/TIC2021S1/T1/crisp_visualguide.pdf) | Acesse aqui <http://meusite.mackenzie.br/rogerio/TIC2021S1/T1/crisp_visualguide.pdf>
* [CRISP DM 1.0](http://meusite.mackenzie.br/rogerio/TIC2021S1$/T1/crisp-dm.pdf) | Acesse aqui <http://meusite.mackenzie.br/rogerio/TIC2021S1$/T1/crisp-dm.pdf>

De qualquer modo, você encontra a seguir, uma descrição resumida de cada uma dessas fases.

## Entendimento do negócio

Esta fase inicial concentra-se em compreender os **objetivos e requisitos do projeto ou problema** de uma perspectiva de negócios, em seguida, convertendo este conhecimento em uma definição de problema de mineração de dados e um plano preliminar projetado para alcançar os objetivos.

Como qualquer projeto é uma fase fundamental para o sucesso do projeto, com o alinhamento das expectativas, definição dos recursos etc.

## Entendimento dos dados

A fase de entendimento dos dados começa com a coleta de dados e segue com atividades que permitem que você se familiarize com os dados, identificar problemas de qualidade de dados, os primeiros insights sobre os dados e / ou detectar subconjuntos interessantes para formar hipóteses sobre informações ocultas.

É uma fase de modelos e hipóteses simples, mas que já fornece muitas respostas relevantes ao problema tratado.

## Preparação de dados

A fase de preparação de dados incluir todas as atividades necessárias para tornar os dados adequados para a modelagem. As tarefas de preparação de dados provavelmente serão realizadas várias vezes e não em qualquer ordem prescrita. As tarefas incluem seleção de dados e atributos, transformações como normalização e recodificação dos dados, limpeza de dados (dados faltantes, ruídos, *outliers* etc.), merge de dados etc.

É importante destacar que, na prática, a atividade de modelagem tem **80% do esforço concentrado nas fases de Entendimento e Preparação dos Dados**. Isto é, não obstante o eventual *glamour* do Cientista de Dados, seu trabalho consiste, na maior parte do tempo, em coletar e manusear grandes conjuntos de dados.

## Modelagem

Nesta fase, várias técnicas de modelagem são selecionadas e aplicadas, e seus parâmetros são calibrados para otimizar valores. Normalmente, existem várias técnicas para o mesmo tipo de problema de mineração de dados. Algumas técnicas têm requisitos na forma de dados. Portanto, voltar à fase de preparação de dados geralmente é necessário. Essa é uma característica importante desse modelo de referência: **as fases não necessariamente seguem sempre em uma única direção**, pois achados em uma etapa posterior podem sempre implicar na necessidade de voltar em tarefas das etapas anteriores.

## Avaliação

Nesta fase do projeto, você construiu um modelo (ou modelos) que parece ter alta qualidade a partir de uma análise de dados perspectiva. Antes de prosseguir para a implantação final do modelo, é importante avaliá-lo completamente e revisar as etapas executadas para criá-lo, para ter certeza de que o modelo atinja adequadamente os objetivos de negócios.

Para isso é necessário **empregar métricas** que permitam avaliar os modelos. Um exemplo conhecido é o empregar o coeficiente de determinação () para avaliação de diferentes modelos de regressão linear, existindo métricas como acuracidade, erro etc. para outros tipos de modelos.

## Distribuição

A criação do modelo geralmente não é o fim do projeto. O resultado do modelo precisa ser organizado e apresentado de forma que possa ser usado. Isso frequentemente envolve a inclusão de modelos dentro dos processos de tomada de decisão de uma organização.

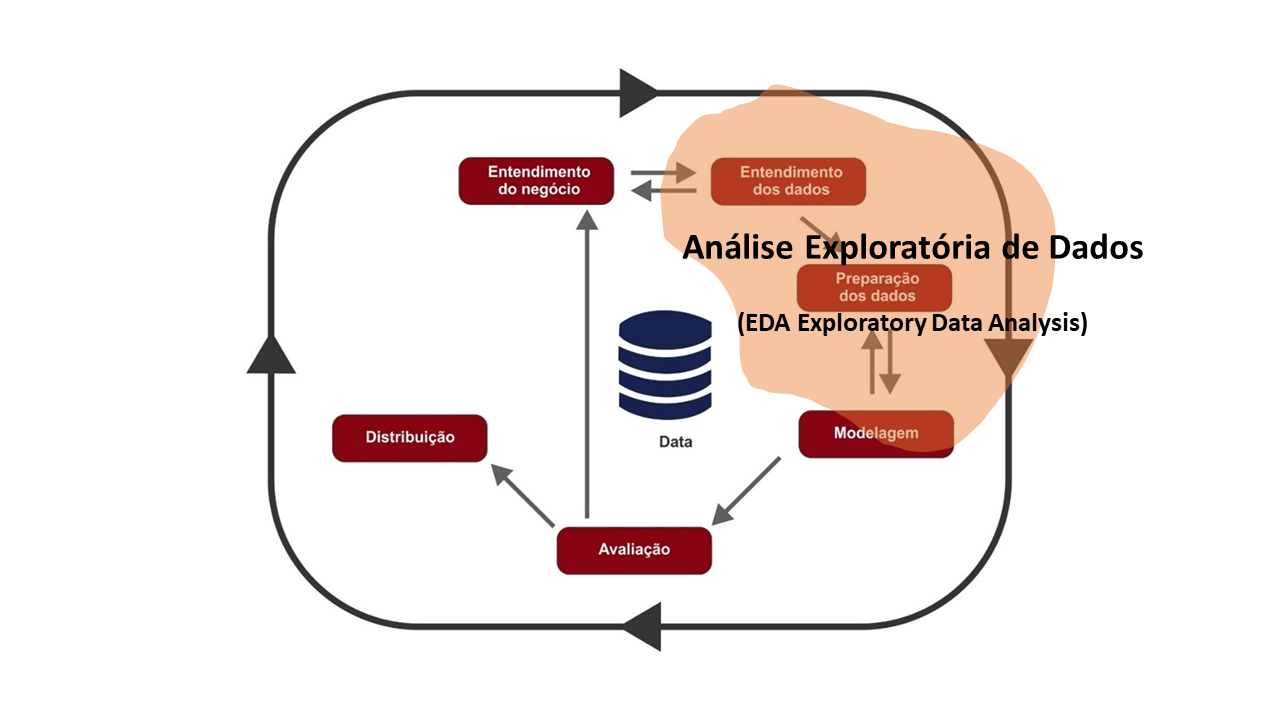
Isso pode envolver uma atividade simples de geração de um relatório em uma página Web até a implantação de diversos artefatos de software para coletar dados de compras de clientes e gerar recomendações automáticas no aplicativo de vendas para outros clientes.

Note que essas fases não são fases estanques e acontecem de forma interativa. Elas muitas vezes são fases que acontecem de modo concorrente com uma fase alimentando a outra. Por exemplo, durante a Preparação dos Dados podemos fazer *descobertas* que não haviam sido feitas durante a fase de Entendimento dos Dados, e mesmo na Modelagem podemos encontrar a necessidade de retornar ao Entendimento dos Dados para verificar algum padrão que não havíamos notado antes.

# Análise Exploratória dos Dados

A Análise Exploratória de Dados (**EDA, Exploratory Data Analysis**) é usada por para analisar e investigar conjuntos de dados e resumir suas características principais, empregando métodos quantitativos e de visualização dos dados.

Dentro do modelo CRISP-DM a EDA compreende as fases de Entendimento e Preparação dos Dados, incluindo também a fase de Modelagem que representamos abaixo de modo parcial para indicar a existência de outros modelos mais elaborados.



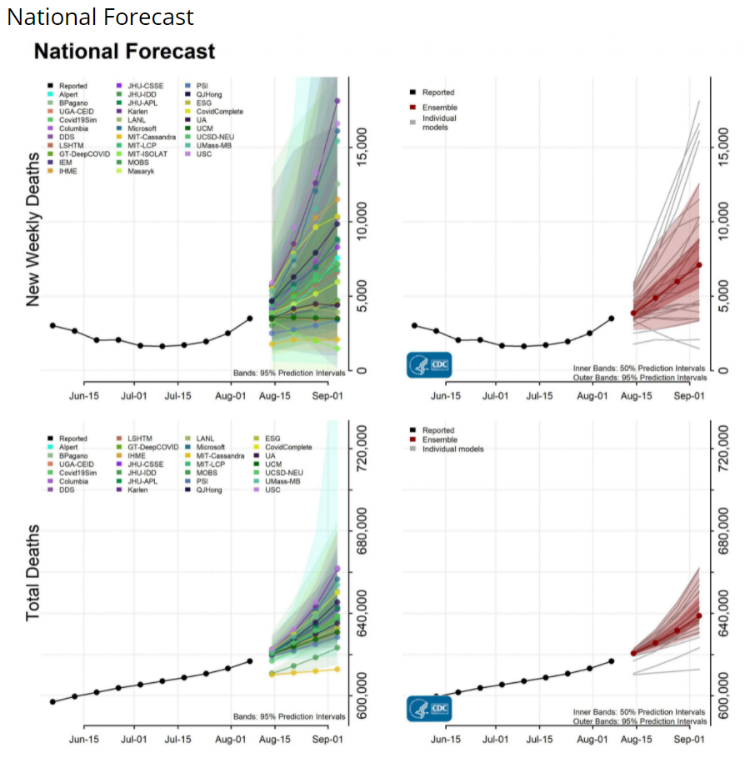
Ela irá permitir responder a perguntas do tipo *O que acontece?* e *Por que Acontece*? trazendo um conhecimento sobre o passado dos dados (veja novamente a figura que relaciona os conceitos de dados, informação e conhecimento).

No nosso exemplo sobre dados de consultas na Web sobre a série *O Gambito da Rainha* você pôde responder que houve um aumento da busca por jogos de Xadrez e eventualmente de produtos relacionados (livros, manuais, jogos online), respondendo à pergunta *O que acontece?*. E pôde também responder à *Por que Acontece*? relacionando a maior procura desses termos ao sucesso da série televisiva!

Este será o foco do nosso curso e iremos instrumentá-lo para que você possa fazer e compreender Análises Exploratórias de Dados (EDA, Exploratory Data Analysis) com diferentes tipos de dados e em diferentes contextos.

## EDA versus Análise Preditiva

Os resultados da EDA por si só já são um grande resultado e permitem tomarmos uma série de decisões baseadas em dados. Mas também podemos querer fazer perguntas direcionadas para o futuro como *O que vai Acontecer?*.



**Evolução e Forecast de Casos de Covid.** (*fonte:* Centers for Disease Control and Prevention, <https://www.cdc.gov/coronavirus/2019-ncov/science/forecasting/forecasting-us-previous.html>)

Responder a esse tipo de pergunta requer não só explorar e entender os dados existentes, mas requer também criar modelos de inferência ou predição. A Análise Exploratória de Dados fornece algumas visualizações e modelos bastante simples, mas respostas mais adequadas são obtidas a partir de modelos preditivos que empregam técnicas de inferência estatística ou aprendizado de máquina e que você estudará em outras disciplinas do curso.

*O que vai Acontecer?*. Você pode pensar nas estimativas de novos casos de Covid empregados para definir ações de política de Saúde Pública ou de um Hospital; nas estimativas das cotações de ações no mercado financeiro ajudando a decidir que ações vender ou comprar; pensar em como um sistema da NetFlix ou do Spotify decide por recomendar um filme ou música que *provavelmente* você terá interesse; ou ainda na detecção de uma *possível* fraude em uma transação de cartão de crédito ou de comércio eletrônico que está para ser realizada.

Mas mesmo para os casos em que buscamos aplicar modelos preditivos, incluindo as aplicações do aprendizado de máquina que hoje está no centro de muitas das soluções de ciência de dados a Análise Exploratória dos Dados é essencial e requerida antes de aplicarmos os modelos preditivos. As perguntas que queremos fazer sobre os dados, raramente são tão claras a ponto de podermos empregar apenas um algoritmo para obtermos as respostas. Antes de aplicarmos o aprendizado de máquina, ou outro método preditivo, pode ser necessário desenvolver métricas, criar representações úteis do problema, fazer hipóteses etc. A EDA (Exploratory Data Analysis) está no centro dessas questões e ajuda a determinar a melhor maneira de manipular as fontes de dados para obter as respostas de que você precisa, tornando mais fácil para os cientistas de dados descobrir padrões, detectar anomalias, testar uma hipótese, e fazer estimativas.

Suponha que você tenha dados do IBGE sobre os brasileiros nos últimos anos, como renda, número de filhos, idade etc. e deseja estimar valores para o próximo ano. Existem muitos métodos simples para isso, mas é importante que o Analista ou Cientista de Dados analise os dados que tem em mãos antes de aplicar modelos preditivos mais poderosos. Por exemplo podem existir muitos dados faltantes (respostas em branco que não foram respondidas ao pesquisador, algumas pessoas omitem a informação de renda) ou você pode ter informações com algum viés (pessoas de mais baixa renda talvez tenham uma propensão maior a informar a renda que pessoas de renda maior) que comprometem a informação para algum grupo de dados. Esses aspectos são então explorados na EDA e são uma parte da análise que precisam preceder aos métodos preditivos ou a aplicação de modelos de aprendizado de máquina.

A EDA permite analisar os dados antes de fazer quaisquer suposições. Ela pode ajudar a identificar erros óbvios, bem como compreender melhor os padrões dentro dos dados, detectar outliers ou eventos anômalos, encontrar relações interessantes entre as variáveis. Os cientistas de dados empregam, então, a análise exploratória para garantir que os resultados que produzem sejam válidos e aplicáveis, antes mesmo de aplicá-los a modelos mais sofisticados de Aprendizado de Máquina.

# Técnicas de Análise Exploratória de Dados

A EDA emprega métodos quantitativos, como estatísticas e sumarizações dos dados, e técnicas de visualização dos dados para obter uma série de análise. Em geral elas envolvem:

* **Análise univariada dos dados** que incluem a análise de cada atributo dos dados individualmente a partir de estatísticas dos dados, análise da distribuição dos valores, identificação de anomalias e padrões etc.
* **Análise bivariadas e multivariadas** que avaliam a relação entre cada variável no conjunto de dados com uma variável objetivo, ou entre várias variáveis.
* **Análise de Agrupamentos**, que incluem a redução de dimensão dos dados em grupos de dados que se assemelham e se diferenciam dos demais

Para essas análises são empregados vários conceitos de Probabilidade e Estatística (estatísticas descritivas dos dados, distribuições de probabilidade, aplicação de probabilidades condicionais etc.) e técnicas de seleção e transformação dos dados (seleção de instâncias e atributos, agregações, transformações de formato, codificações de valores, criação de atributos derivados, merge de dados etc.). Várias desses conceitos e técnicas você aprenderá e aplicará aqui criando Análises de Dados com R e Python.

# Linguagens e Ferramentas de Ciência de Dados

Há uma ampla discussão sobre qual é a melhor linguagem para empregarmos em Ciência de Dados. Mas essa é uma discussão que parece mais próxima de *qual é o melhor vinho?*, e a melhor resposta para essa pergunta parece ser *é o que você gosta*. Alguns critérios ajudam a entender por que Python & R são melhores linguagens que outras para a aplicação em Ciência de Dados: são ambas de código aberto; apresentam uma menor curva de aprendizado; apresentam ótima velocidade de processamento; proporcionam grande produtividade de desenvolvimento (particularmente por serem linguagens de scripts); possuem amplas capacidades de tratamento e visualização de dados (em parte obtidas por pacotes adicionais às linguagens); possuem gnrade confiabilidade do código; e têm amplo uso no mercado e comunidades; Todas essas características tornam o Python e o R linguagens mais adequadas para o emprego em Ciência de Dados que outras linguagens de propósito geral (C, C++, Java, PHP, Pearl etc.) e as tornam bastante similares.

Para dizer um mínimo das diferenças, R parece apresentar capacidades gráficas e de visualização de dados superiores, além de uma curva de aprendizado menor para não programadores; O Python, entretanto, vem sendo uma linguagem extremamente popular e grandes desenvolvimentos no aprendizado de máquina na última década, como *deep learning*, são desenvolvidos em Pyhton. Também há uma diferença de recursos disponíveis para essas linguagens dependendo do campo de aplicação. A áreas da saúde e biologia, ciências sociais, e áreas de economia e finanças há anos empregam o R. Já áreas como comércio eletrônico, empresas digitais e soluções mais ligadas a automação e Inteligência Artificial, dão preferência ao uso de Python. Assim, o mais correto parece decidir a *"melhor linguagem"* de acordo com a organização e propósito que será empregado. Aqui iniciamos nosso curso com R e deixaremos para fazer Análises de Dados com Python mais adiante, depois de introduzirmos os conceitos necessários.

*Ainda existem ferramentas do tipo* ***spreadsheet view of data****, ferramentas que acessam os dados em planilhas que podem ser imediatamente manipuladas pelos usuários (como Tableau, Qlik e outros) e ferramentas do tipo* ***drag-and-drop****, em que funções pré definidas podem ser arrastadas e coladas para compor um fluxo de processamento (como disponível em ferramentas como o SAS, SAS VA, Tableau, RapidMiner, Weka). Isso tem suas vantagens e desvantagens. Ferramentas visuais buscam aumentar a produtividade, mas o uso de linguagens de programação (código) em geral dá maior flexibilidade e permite maior reuso das soluções. Assim como no caso anterior o uso de uma solução ou outra deve ser analisada caso a caso, de acordo com a organização e propósito para que será empregada, sem necessariamente assumirmos que uma ou outra é a melhor solução.*

# Ambientes R

É esperado que você nesse ponto já tenha alguma familiaridade com alguma linguagem de programa (Python, C++, Java) e editores de programação. Aqui você aprender como empregar um ambiente R para criar e executar programas e a partir da próxima aula faremos uma introdução à linguagem baseada em exemplos, assumindo que você já tem os conceitos básicos como variáveis, tipos de dados, atribuição, controles de fluxo em programas etc.

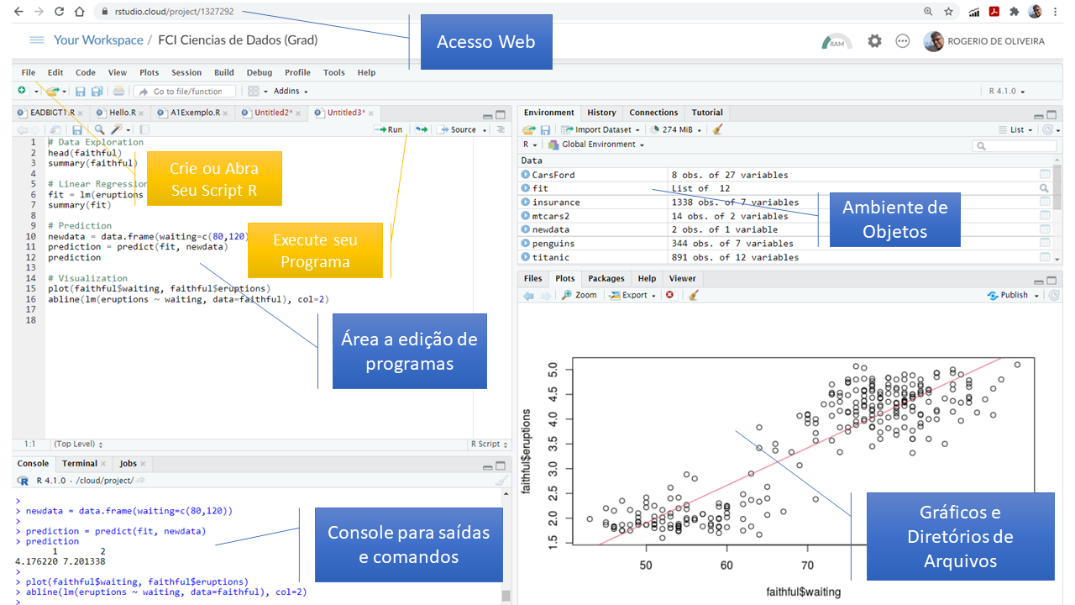
## Acessando o RStudio

Para programar em R você pode empregar o ambiente R Studio, o [RStudio Cloud](https://rstudio.cloud/) (<https://rstudio.cloud/>) é um ambiente em nuvem, gratuíto, para o desenvolvimento de aplicações. Mas se prefirir o uso de um ambiente local acesse [R Project](https://cran.r-project.org/mirrors.html)(<https://cran.r-project.org/mirrors.html>) e [RStudio](https://rstudio.com/products/rstudio/download/#download)(<https://rstudio.com/products/rstudio/download/#download>) fazendo a instalação do R e depois do RStudio. Os ambientes e interface são bastante semelhantes.

Como outros editores de linguagem de programação você encontra no RStudio uma barra de opções (arquivo, edit, code, view, ...) e áreas de trabalho:

* Uma área para edição dos programas
* Uma console para comandos e saídas dos programas
* Uma área de arquivos, plots etc.
* Uma área de objetos, onde você pode inspecionar o estado de suas variáveis

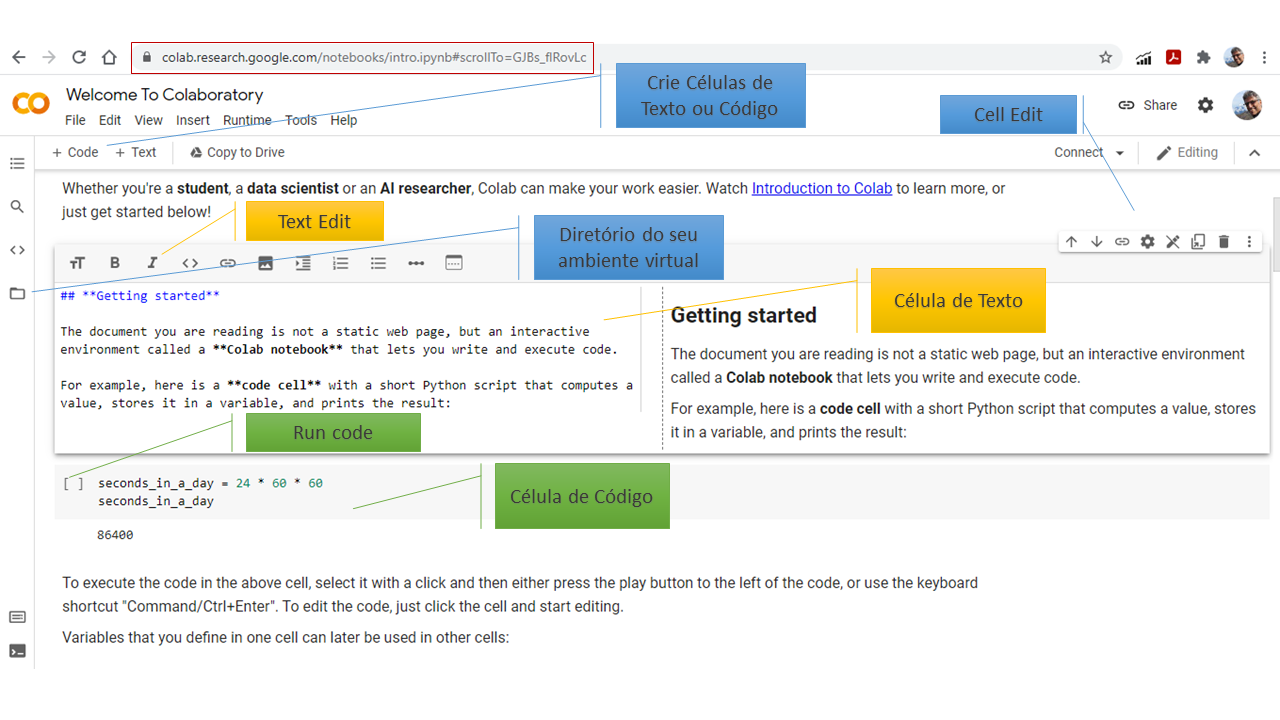
Que apresentamos abaixo.



Essa pode ser uma opção interessante quando você quiser produtividade ou criar múltiplos programas que interagem entre si. Também permite criar documentos R Markdown e R Notebooks que permitem a criação em um único arquivo código, saídas e texto, muito semelhante aos notebooks Python que você já deve ter conhecido.

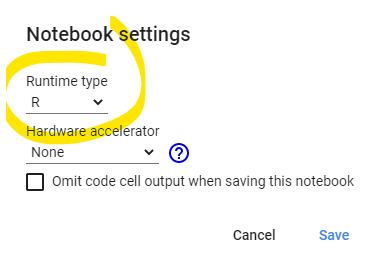
## Ambiente Google Colaboratory ou Colab para R

A outra opção, e a que empregaremos ao longo do curso é o uso do [Google Colaboratory](https://colab.research.google.com/) ou simplesmente *Colab*. Você jé deve ter conhecido esse ambiente para a programação Python. Este é um ambiente que permite você escrever código diretamente em seu navegador, sem nenhuma configuração ou instalação de softwares necessária, e que ainda fornece acesso a recursos de GPU para o processamento e compartilhamento fácil de código (GitHub, Google Drive etc.). Para acessar, basta digitar no seu navegador <https://colab.research.google.com/>.



Mas este mesmo ambiente pode ser igualmente empregado para R. Por permitir ainda a inclusão em um mesmo documento código, saída das execuções e textos, ele é um ambiente ideal para o aprendizado e vem sendo muito empregado para criação de Análises de Dados.

A execução de notebooks R no Colab, entretanto, requer que você inicie o Colab com <http://colab.to/r> e então abra os arquivos .ipynb com código R. Você pode verificar se a sua sessão Colab está setada para executar programas R entrando na barra de ferramentas em Runtime > Change runtime type, na qual deverá aparecer a opção R como abaixo.



Esse será nosso ambiente preferencial de programação por permitir empregar tanto R como Python.

O script abaixo é um programa R e você pode copiar e executar este código em quaisquer dos ambientes acima para apenas explorar o ambiente de execução do R. Sua familiaridade com o ambiente será muito útil para que você possa aprender a partir da próxima aula os conceitos e a programação R.

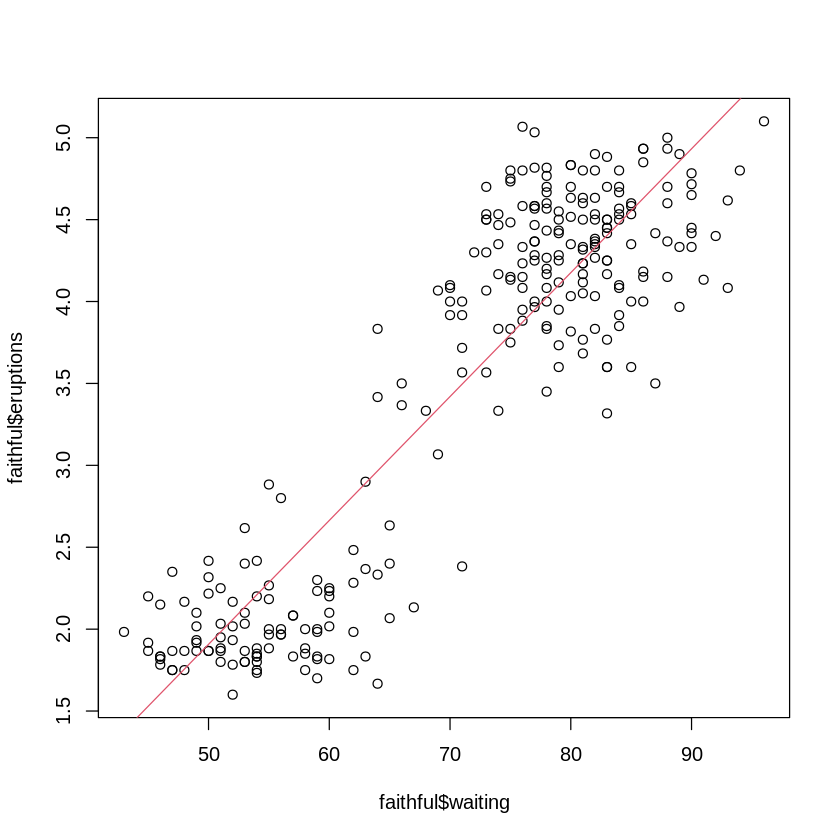
# Data Exploration   
head(faithful)  
summary(faithful)  
  
# Linear Regression  
fit = lm(eruptions ~ waiting, data=faithful)  
summary(fit)  
  
# Prediction  
newdata = data.frame(waiting=c(80,120))  
prediction = predict(fit, newdata)   
prediction  
  
# Visualization  
plot(faithful$waiting, faithful$eruptions)  
abline(lm(eruptions ~ waiting, data=faithful), col=2)

eruptions waiting  
1 3.600 79   
2 1.800 54   
3 3.333 74   
4 2.283 62   
5 4.533 85   
6 2.883 55

eruptions waiting   
 Min. :1.600 Min. :43.0   
 1st Qu.:2.163 1st Qu.:58.0   
 Median :4.000 Median :76.0   
 Mean :3.488 Mean :70.9   
 3rd Qu.:4.454 3rd Qu.:82.0   
 Max. :5.100 Max. :96.0

Call:  
lm(formula = eruptions ~ waiting, data = faithful)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-1.29917 -0.37689 0.03508 0.34909 1.19329   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) -1.874016 0.160143 -11.70 <2e-16 \*\*\*  
waiting 0.075628 0.002219 34.09 <2e-16 \*\*\*  
---  
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  
  
Residual standard error: 0.4965 on 270 degrees of freedom  
Multiple R-squared: 0.8115, Adjusted R-squared: 0.8108   
F-statistic: 1162 on 1 and 270 DF, p-value: < 2.2e-16

1 2   
4.176220 7.201338

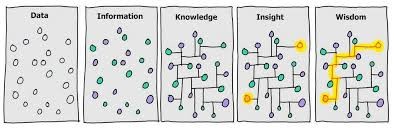


O programa acima emprega um conjunto de dados sobre erupções. Ele apresenta estatísticas sumarizadas dos dados (eruptions e waiting) para, em seguida criar um modelo linear simples para os dados:

O modelo é representado no gráfico e a aderência do modelo aos dados é avaliada mostrando um coeficiente de determinação (Adjusted R-squared) de , o que significa que o modelo explica aproximadamente % da variação dos dados. Você não precisa compreender este código agora. Ele está aqui apenas para ilustrar a criação de Análises de Dados em R e para que você tenha um script R que possa executar e experimentar os ambientes R.

# Para Saber Mais

1. A figura abaixo representa de modo bastante mais informal as relações entre os conceitos de dados, informação e conhecimento que a figura que empregamos antes. Por outro lado, ela parece ajudar muito no entendimento desses conceitos e dá ênfase na conexão entre os dados. Examine a figura abaixo e reflita sobre como ela representa as relações entre os conceitos de dados, informação e conhecimento associando exemplos para cada uma das categorias apresentadas. Compartilhe com seus colegas os exemplos e suas opiniões.



1. Leia sobre algumas das diferenças entre as Linguagens R e Python para Ciência de Dados em <https://www.ibm.com/cloud/blog/python-vs-r>. Em seguida pesquise na internet aplicações de R e Python em Ciência de Dados.
2. Vamos fazer uma Análise Exploratória de Dados com o [Google Trends](https://trends.google.com.br/)? Acesse (<https://trends.google.com.br/>) procure identificar produtos ou serviços que tiveram a sua procura alterada a partir do início da Pandemia no Brasil (a partir do início de 2020). Busque por produtos ou serviços que passaram a ter menos procura e pelos que tiveram aumento. Compartilhe então seus resultados.
3. Dados públicos e externos são cada mais empregados para análises de dados. Existem vários sites em que você pode pesquisar sobre indicadores financeiros como IBOVESPA ou a taxa do dólar e sobre a evolução dos casos de Covid a partir do início da Pandemia no Brasil em meados de 2020. Pesquise e responda, qual o impacto dos primeiros casos de Covid no mercado de ações brasileiro (IBOVESPA) e no preço do dólar? Empregue visualizações desses dados para visualizar e apenas estimar aproximadamente esses impactos. Compartilhe suas conclusões indicando as referências (sites) empregados.

# Referências

1. Foster Provost, Tom Fawcett. **Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking**.
2. Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, Rüdiger Wirth [**CRISP DM 1.0**](http://meusite.mackenzie.br/rogerio/TIC2021S1$/T1/crisp-dm.pdf) **Step-by-step data mining guide** Disponível em: <http://meusite.mackenzie.br/rogerio/TIC2021S1$/T1/crisp-dm.pdf> Acesso em: 29 de Agosto de 2021.
3. Hadley Wickham, Garrett Grolemund. **R for Data Science: Import, Tidy, Transform, Visualize, and Model Data.** O'Reilly Media Inc. (2017). (alternativamente disponível em: <https://r4ds.had.co.nz/>).