# Para que serve?

Data Science ou Ciência de Dados é "um campo de estudo interdisciplinar que faz uso de métodos científicos para extrair padrões, insights, conclusões e previsões a partir de dados, os quais podem ser dados tabulares, texto, imagens, vídeo, som, entre os mais diversos tipos. Nesse contexto, um Cientista de Dados é uma pessoa que aplica tais métodos, muitas vezes usando técnicas de Big Data."

No entanto, existem áreas adjacentes que merecem atenção adequada e não devem ser confundidas com Ciência de Dados.

Um deles é a Ciência da Decisão. Sua importância não deve ser subestimada, por isso é útil conhecer as reais diferenças e peculiaridades desses dois campos.

Ciência de Dados e Ciência da Decisão são campos relacionados, mas ainda separados, portanto, em alguns pontos, pode ser difícil compará-los diretamente.

Conforme mencionado, um Cientista de Dados é um especialista envolvido na busca de insights de dados após estes dados serem coletados, processados e estruturado pelo Engenheiro de Dados. O Cientista da Decisão considera os dados como uma ferramenta para tomar decisões e resolver problemas de negócios.

Para demonstrar outras diferenças, vamos dar uma olhada no infográfico abaixo que coloca a Ciência de Dados e Ciência da Decisão em contraste de acordo com vários critérios



Em termos de definição, a Ciência de Dados é um campo interdisciplinar que usa algoritmos científicos, métodos, técnicas e várias abordagens para extrair percepções valiosas. Seu objetivo principal é revelar as percepções dos dados para posterior aplicação em benefício dos vários setores.

Em contraste, a Ciência da Decisão é uma aplicação de um complexo de técnicas quantitativas ao processo de tomada de decisão. Seu objetivo é aplicar os insights baseados em dados em combinação com os elementos da ciência cognitiva para o planejamento e desenvolvimento de políticas. Portanto, os dados são igualmente importantes para ambos, mas os mecanismos são diferentes.

A Ciência de Dados é aplicada em vários setores, como varejo, entretenimento, mídia, saúde, seguros, telecomunicações, finanças, viagens, manufatura, agricultura, esportes, etc. A Ciência da Decisão é aplicada mais em áreas teóricas de negócios e gestão, direito e educação, meio ambiente regulamentação, ciência militar, saúde pública e políticas públicas.

A Ciência de Dados pode ser um componente crucial da Ciência da Decisão e, muitas vezes, os proprietários de negócios contam com a Ciência de Dados como uma solução para todos os seus problemas e preocupações. No entanto, não basta usar apenas a Ciência de Dados. A verdade está em algum lugar entre a Ciência de Dados e a Ciência da Decisão.

# Arquitetura de Dados

A arquitetura de dados é um daqueles componentes invisíveis, mas que, quando falham, podem comprometer seriamente a performance de uma empresa e seus profissionais.

Arquitetura dos dados é a forma que diferentes empresas encontram para organizar seus dados, ativos digitais e as relações entre eles.

Ou seja, assim como organizamos nossas roupas em armários, closets e gavetas de diferentes maneiras, o mesmo também é feito com a informação em formato eletrônico.

Por isso, uma arquitetura eficiente é aquela que garante o acesso aos dados com segurança, no tempo adequado e de forma inteligível para seus usuários.

Todos os profissionais de TI e empresas que lidam com soluções analíticas sabem que uma arquitetura de base de dados só traz resultados quando se orienta por seis princípios, sendo eles:

* SEGURANÇA

Pelos bons princípios de governança de dados, todo sistema deve ser seguro o bastante para que as informações em uma companhia sejam acessíveis apenas às pessoas autorizadas. Dessa forma, a arquitetura deve observar mecanismos de proteção contra acessos indevidos ou invasores, franqueando os dados críticos somente àqueles que forem credenciados para isso.

* FLEXIBILIDADE

Por outro lado, um sistema seguro não quer dizer que ele não possa ser moldado conforme as necessidades. Afinal, na transformação digital, é preciso que a arquitetura dos dados tenha certa elasticidade, permitindo que os sistemas evoluam e sejam escaláveis. Há casos, ainda, em que é necessário conceder novos acessos ou autorizações de uso não previstas. Então, quanto mais ela antecipar essas demandas, melhor.

* COLABORAÇÃO

O modelo de gestão horizontal é cada vez mais uma tendência.Por isso, as empresas precisam de soluções que permitam gerir, acessar e tratar seus dados por múltiplos times e colaboradores.

* INTELIGÊNCIA

Assim como há hoje edifícios inteligentes, o mesmo se espera dos sistemas pelos quais as companhias lidam com informações estratégicas. Esse é o princípio por trás do conceito de Business Intelligence (BI), pelo qual as atividades são pautadas por decisões tomadas a partir de dados estruturados. Logo, é tarefa dos profissionais de arquitetura de banco de dados garantir que a empresa terá à sua disposição não apenas dados em estado bruto, mas informação útil sempre que precisar.

* AUTOMAÇÃO

Não dá para imaginar uma ferramenta digital que abra mão de processos automatizados. Por esse princípio, a arquitetura dos dados assume o compromisso de gerar soluções que sejam efetivas e em níveis máximos de automação.

* ORIENTADA PARA RESULTADOS

Na teoria, a arquitetura de banco de dados parece ser perfeita e solucionar todos os problemas relativos ao uso da tecnologia. No entanto, ela só será útil na prática quando for orientada pelas metas do negócio.

Nós podemos construir uma casa habitável, mas mal projetada, fazendo com que a luz do sol se direcione sobre os cômodos na pior hora possível.

O mesmo se aplica à Arquitetura de Dados, onde um projeto mal feito pode não só dificultar o acesso aos dados como comprometê-los de forma difícil de contornar.

## Data Lake

Um Data Lake é um repositório de armazenamento que pode armazenar grande quantidade de dados estruturados, semiestruturados e não estruturados.

É um local para armazenar todos os tipos de dados em seu formato nativo, sem limites fixos no tamanho ou no arquivo da conta. Oferece alta quantidade de dados para aumentar o desempenho analítico e a integração nativa.

## Big Data

“Data is the new science. Big Data holds the answers.” - Patrick P. Gelsinger

Antes de tudo, Big Data também são dados, mas com um tamanho muito maior. Big Data é um termo usado para descrever uma coleção de dados com volume enorme e que cresce exponencialmente, a todo momento.

Em resumo, esses dados são tão grandes e complexos que nenhuma das ferramentas tradicionais de gerenciamento de dados é capaz de armazená-los ou processá-los com eficiência.

### Os 'Vs' do Big Data

Em 2001, Gartner articulou a definição do termo Big Data em três 'Vs': Volume, Velocidade e Variedade. Os 'Vs' nos ajudam a ter uma compreensão melhor de como funciona Big Data, apesar de ser antigo.

* Volume

O volume dos dados é importante. Com o Big Data, você terá que processar grandes volumes de dados não estruturados e de baixa densidade.

Para algumas organizações, isso pode ser dezenas de terabytes de dados. Para outras, pode ser centenas de petabytes.

* Velocidade

Velocidade é a taxa mais rápida na qual os dados são recebidos. Normalmente, a velocidade mais alta dos fluxos de dados são feitas diretamente na memória, ao contrário da gravação em disco.

* Variedade

Variedade refere-se aos muitos tipos de dados disponíveis. Os tipos de dados tradicionais foram estruturados e se encaixam perfeitamente em um banco de dados relacional. Com o aumento do Big Data, os dados começaram a surgir em novos tipos, como os não estruturados.

Com o passar do tempo e da quantidade de dados produzida aumentando drasticamente, a IBM introduziu mais dois Vs, que seria Veracidade e mais importante do Big Data, o Valor.

* Veracidade

A Veracidade refere-se à qualidade dos dados que estão sendo analisados. Os dados de alta veracidade têm muitos registros valiosos para analisar e que contribuem de maneira significativa para os resultados gerais. Os dados de baixa veracidade, por outro lado, contêm uma alta porcentagem de dados sem sentido.

* Valor

O valor que os dados geram para os usuários e para os negócios. Já sabemos que temos um volume colossal de dados gerados a cada minuto de diversas fontes e formatos, dados estes que devem ser verídicos para gerar valor.

# Análise de dados

A Análise de Dados é a arte de transformar dados em conhecimentos e insights relevantes. Ou seja, comparar ou agregar as informações brutas para entender o que os dados nos dizem.

**

Mais informações no blog [O Guia Completo Sobre Data Science: Tudo O Que Você Precisa Saber!](https://letscode.com.br/blog/o-guia-completo-sobre-data-science-tudo-o-que-voce-precisa-sabe)

# Metodologia de Análise de Dados

* Definição das perguntas

A primeira etapa da Metodologia de Análise de Dados diz respeito à definição das perguntas que você pretende responder ao analisar os dados que você tem em mão.

As perguntas devem ser mensuráveis, claras e concisas. Elabore suas perguntas para qualificar ou desqualificar possíveis soluções para seu problema ou oportunidade específica.

Imagine a seguinte situação. A sua empresa tem enfrentado problemas com um fornecedor; o contratado está gerando custos crescentes para você e tem se mostrado incapaz de oferecer propostas competitivas no contrato.

Nesse caso, uma pergunta que você pode fazer para solucionar esse problema é:

A empresa tem condições de arcar com uma multa de quebra de contrato e contratar outro fornecedor que ofereça melhores condições?

* Defina o que medir

Usando o exemplo do fornecedor citado anteriormente, considere que tipo de dados você precisaria para responder à sua pergunta-chave.

Nesse caso, você necessita saber, por exemplo, os custos que aquele fornecedor específico representa para a sua empresa.

Depois, certifique-se de incluir toda e qualquer objeção razoável ​​que tanto a sua empresa quanto o fornecedor possa apresentar. Se você for, de fato, trocar de fornecedor, procure quantificar os impactos dessa decisão.

* Defina como medir

Nesta terceira etapa da Metodologia de Análise de Dados, é muito importante pensar sobre como você mede seus dados. Isso porque o processo de mensuração pode afetar a qualidade dos dados coletados.

Veja abaixo algumas perguntas que podem ser feitas para essa etapa:

Qual é o seu prazo disponível para analisar os dados?

Qual será a fonte dos dados?

Quais fatores devem ser incluídos na análise?

Levando isso em conta, estabeleça critérios que permitirão coletar os dados em tempo hábil, sem afetar sua qualidade.

* Faça a coleta de dados

Com suas perguntas e suas prioridades de medição claramente definidas, é hora de coletar seus dados.

Nesta etapa do Metodologia de Análise de Dados, determine quais informações podem ser coletadas de bancos de dados ou fontes existentes. Colete esses dados primeiro.

Determine com antecedência um sistema de armazenamento e nomeação de arquivos para ajudar todos os membros da sua equipe a colaborar. Esse processo economiza tempo e impede que os membros da equipe coletem as mesmas informações duas ou mais vezes.

Se você precisar coletar dados por meio de observação ou entrevistas, desenvolva um modelo de questionário com antecedência a fim de otimizar o tempo. Mantenha seus dados coletados devidamente identificados e organizados.

* Faça a análise de dados

Depois de coletar os dados certos para responder à pergunta que você fez na 1ª Etapa, chega-se à análise aprofundada desses dados.

Você pode recorrer a gráficos, tabelas e outros recursos visuais que permitam uma visualização mais clara dos dados coletados.

Uma tabela dinâmica, por exemplo, pode te ajudar a classificar e filtrar os dados a partir de diferentes variáveis. Durante esta etapa do nosso método de análise de dados, a utilização de softwares pode ser extremamente benéfica.

À medida que você for manipulando os dados, é possível que você precise revisar sua pergunta original ou coletar mais dados. De qualquer forma, essa análise inicial ajuda você a concentrar sua análise de dados para responder melhor à sua pergunta e a quaisquer objeções que possam aparecer.

* Interprete os resultados

Depois de analisar seus dados e possivelmente realizar mais pesquisas, finalmente é hora de interpretar seus resultados. Ao interpretar sua análise, faça perguntas como:

Os dados respondem à pergunta original? Se sim, de que forma?

Os dados ajudam você a se defender de qualquer objeção? Se sim, de que forma?

Existe alguma limitação nas suas conclusões? Se sim, quais?

Se a sua interpretação dos dados se sustentar sob todas essas questões e considerações, provavelmente você chegou a uma conclusão produtiva.

### Podemos resumir as etapas de um processo de Análise de Dados em 4 pontos principais sendo eles:

* Definição dos problema a ser resolvido

Parte mais importante do processo de análise de dados! Aqui o objetivo é entender “O QUÊ” a análise vai resolver, ou estudar.

* Preparação e Exploração dos dados

É aqui onde definimos “COMO” resolver o problema. Uma vez que sabe qual o problema a ser resolvido podemos traçar um plano de ação.

(Coleta de dados, Preparação e transformação dos dados, Limpeza dos dados)

* Análise Exploratória de dados

A Analise exploratória dos dados é considerado um dos passos cruciais para as analises em base de dados. Isto quer dizer, que existirá um processo para organizar, resumir, aplicar alguns cálculos e visualizar os dados presentes na sua analise. Desta forma, resumindo as principais características por meio de métodos visuais.

Este processo foi desenvolvido por um dos maiores estatísticos de todos os tempos, John W. Tukey. Com o livro Exploratory Data Analysis (EDA) lançado em 1977, fez uma contribuição memorável para o mundo cientifico e social, unindo pensamento estatístico aos processos de transformação e exploração de dados.

Portanto, na Analise exploratória dos dados identificamos os comportamentos médios e discrepantes, procurando e identificando tendências, avaliando os comportamentos e investigando a interdependência entre as variáveis. Assim, podemos entender que as ferramentas para prover a Analise exploratória dos dados são fundamentais, logo, não abra mão de utilizar todas as técnicas adequadas e disponíveis para a sua Analise exploratória dos dados.

(Tabelas de dados, Histogramas, Box Plot, Scatter Plot, Five-Number Summary, Analise de Outliers)

* Criação do Modelo

Um modelo nada mais é do que é uma função matemática\estatística para entender, ou até mesmo prever situações com base nos dados.

* Apresentação dos resultados

A principal tarefa nesta etapa do processo de Análise de Dados é transformar informações complexas em simples, é traduzir e transmitir os resultados obtidos através das pesquisa da forma mais clara possível.

Aqui se aplica a habilidade de storytelling e comunicação. A pessoa que será responsável pela tomada de decisão, com base nos dados da pesquisa, precisa entender perfeitamente o significado dos resultados e conclusões que a pesquisa obteve.

# Métricas e KPIs

Você mensura as ações realizadas na sua empresa? Se sim, provavelmente utiliza os indicadores métrica e KPI, certo? Mas antes de mais nada, você sabe o que são e qual é a diferença entre eles?

Os resultados das métricas e KPIs podem trazer muitas respostas para o planejamento estratégico de uma organização e é muito importante saber para o que eles servem e como podem ser utilizados.

### O que são métricas?

As métricas possibilitam mensurar, monitorar e gerir as estratégias de uma empresa. Elas apresentam informações sobre quais estratégias devem ser continuadas, aperfeiçoadas ou abandonadas.

### O que são KPIs?

KPIs ou Indicadores chave de desempenho, são criados a partir das métricas. São eles que indicam de forma percentual ou numérica o resultado de uma ação e por onde é avaliado o desempenho da organização.

Eles são essenciais nos diversos departamento de uma empresa, pois apontam os reais resultados dos investimentos feitos, seja em gestão de pessoas, na otimização de processos, na redução de gastos, no aumento da produtividade dos colaboradores e vários outros.

Entre as categorias existentes, algumas bastante utilizadas são:

* Indicadores de produtividade

Esse KPI se refere ao esforço dos funcionários x resultados entregues. Ou seja, quanto foi produzido pelo funcionário e quais recursos foram necessários para alcançar este resultado.

* Indicadores de capacidade

Indica qual a capacidade de produção e em quanto tempo.

Ex: Quantas peças uma fábrica consegue produzir em um determinado período.

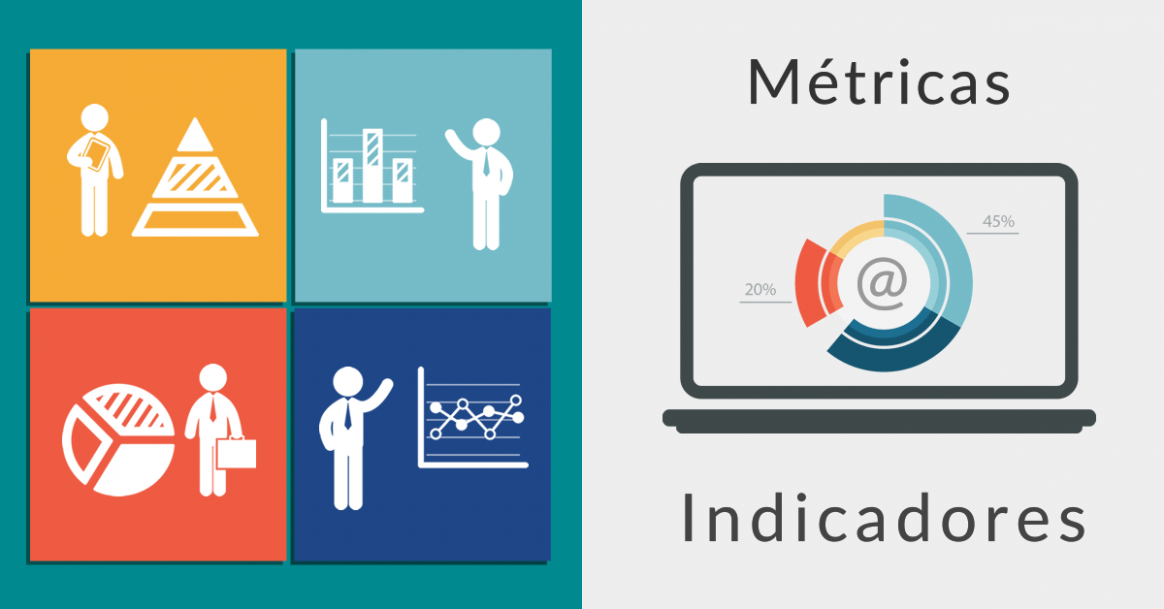
* Indicadores de qualidade

Está ligado a satisfação e aceitação dos clientes com um produto ou serviço entregue. Se o produto ou serviço foi entregue dentro de todos os parâmetros de qualidade estabelecidos pela empresa.

* Indicadores estratégicos

Aponta a situação da empresa em relação ao planejamento feito. Possibilita comparar resultados, controlar objetivos e repensar formas para melhoria contínua de processos.

### Diferenças entre métrica e KPI



Métrica e KPI são indicadores de gestão, mas basicamente a diferença entre eles é que um é a base do outro, um dá continuidade ao outro.

As métricas são informações sobre o negócio, os KPIs são resultados de alguma ação que foi tomada com base nas métricas.

# Construindo Dashboards Analíticos

## Objetivo

Nesta seção veremos como construir Dashboards Analiticos com foco em tranformar os dados armazenados em informação para tomada de decisão em nossas empresas.

## Recursos para a Aula

* PostgreSQL
* PowerBI Desktop

Agora que já sabemos a construir querys e objetos com SQL vamos aprender a entregar valor para o négocio construindo Dashboards que transformem nossos dados armazenados nas tabelas de nossos bancos de dados em informação de alto valor para tomada de decisão em nossas empresas.

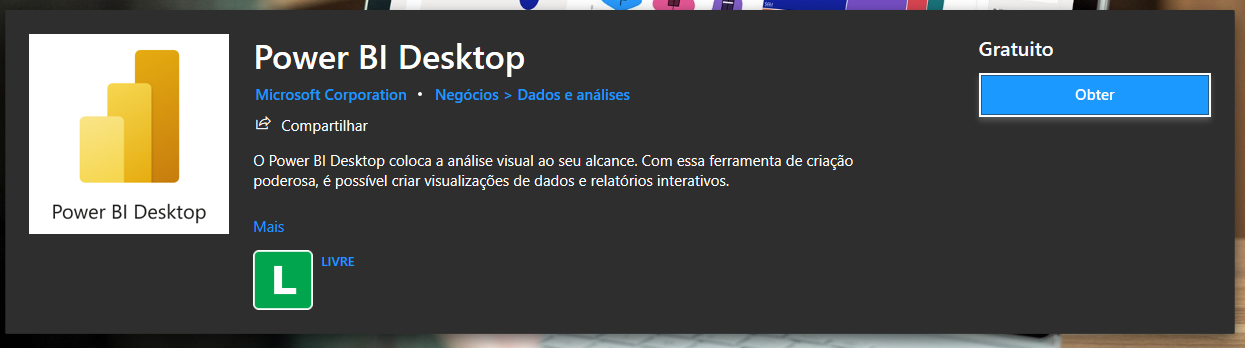
Para que isso seja possivel precisamos de um software que transforme os dados armazenados em nossos bancos de dados em gráficos de fácil compreensão para que as pessoas que possuam conhecimento do negócio passam interpreta-los de forma analítica.

Existem diversos softwares que podemos utilizar para tal atividades, dentre eles temos:

* Tableau
* Microsoft Power BI
* MoData
* Qlick Sense
* Looker

Nesta aula iremos utilizar o Microsoft Power BI. Esta é uma ferramenta que nasceu dentro do Excel e desde que foi separada deste vem crescendo e sendo aperfeiçoada.

Para instalar basta [clicar no link](https://powerbi.microsoft.com/pt-br/downloads/), fazer o download do Microsoft Power BI Desktop e fazer a instalação diretamente (Next-Next-Finish):



## Dashboards usando Microsoft Power BI

Vamos construir um dashboard baseado em nosso banco de dados Northwind.

O banco de dados Northwind é um banco de dados de exemplo usado pela Microsoft para demonstrar o recursos de alguns de seus produtos, incluindo SQL Server e Microsoft Access.

O banco de dados contém os dados de vendas da Northwind Traders, uma importação e exportação de alimentos especiais fictícios empresa.

### Preparando o ambiente...

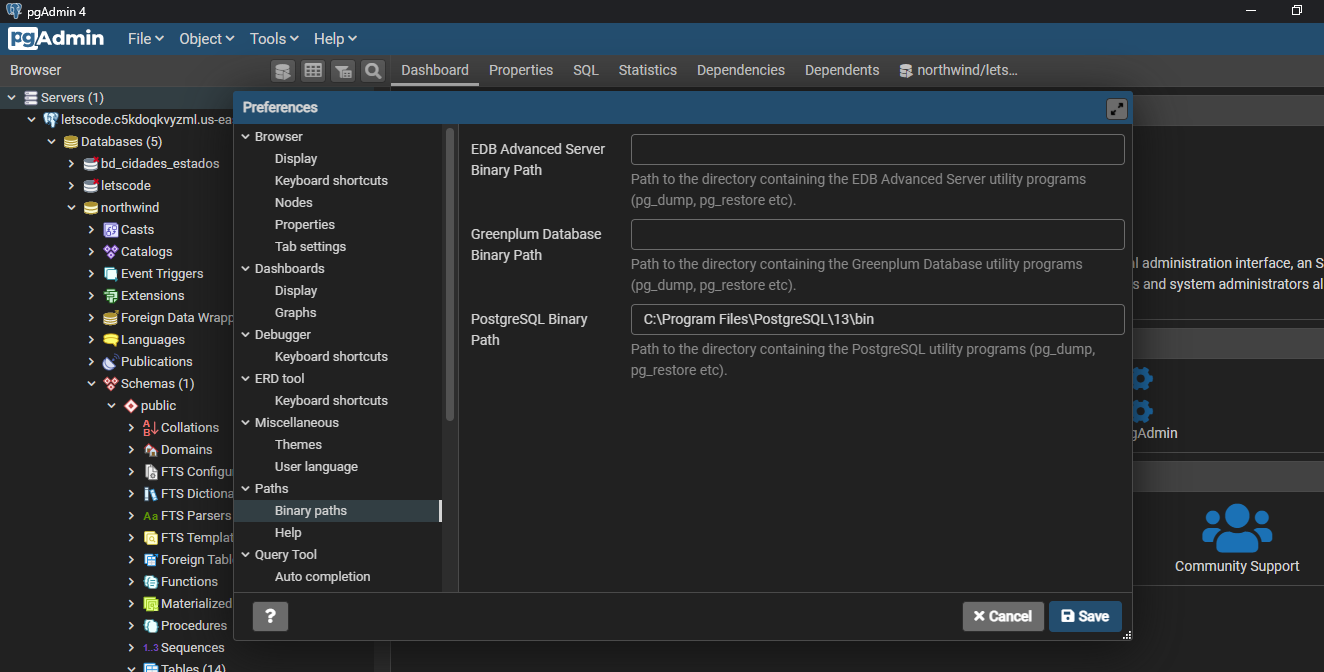
* Conectado ao Banco de Dados RDS Postgres, crie o database northwind.

/\*Estrutura do database northwind\*/

CREATE DATABASE northwind WITH OWNER = letscode ENCODING = 'UTF8' CONNECTION LIMIT = -1;

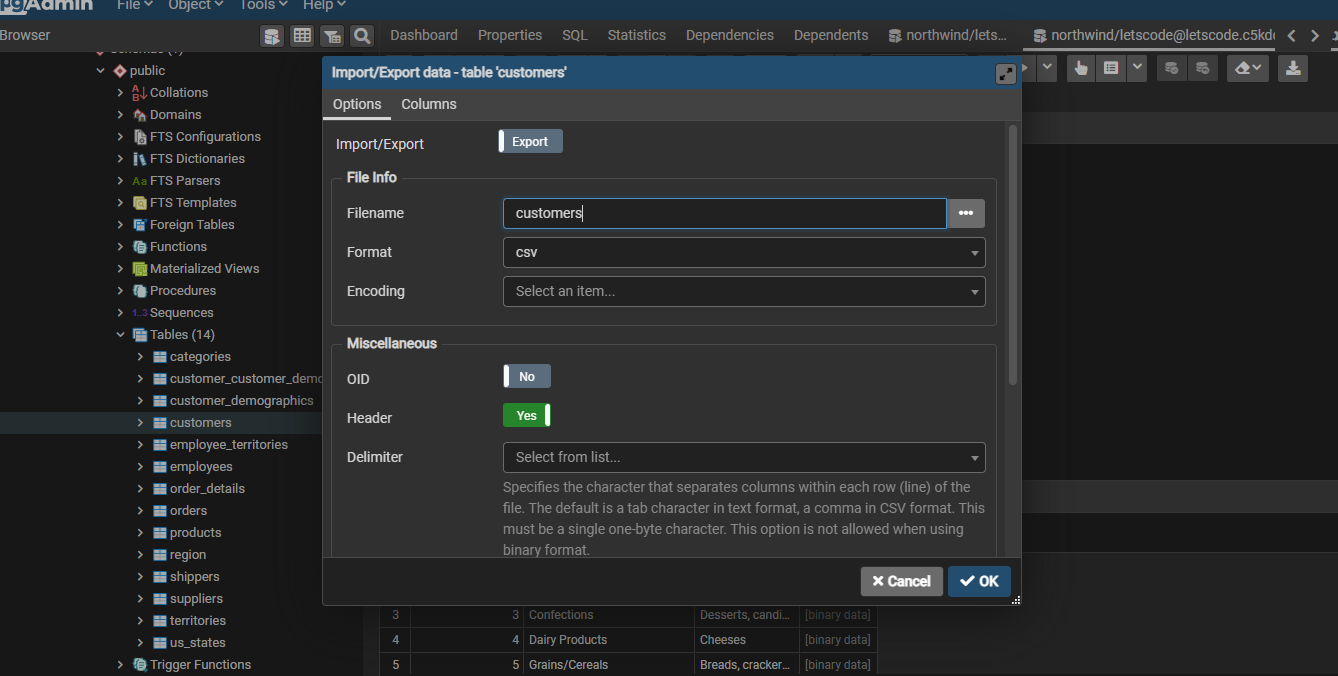
~

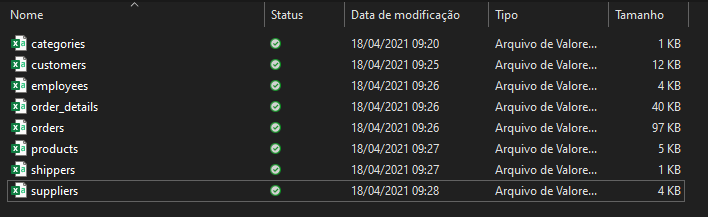
* Vamos utilizar o arquivo [northwind.sql](https://s3-sa-east-1.amazonaws.com/lcpi/07d483a2-9c46-4969-8ab9-e45e2e8d2ac3.sql) para criar as tabelas Customers, Categories, Employees, Order\_Details, Orders, Products, Shippers, Suppliers e inserir registros nessas tabelas.
* Configurar o PostgreSQL Binary Path

****

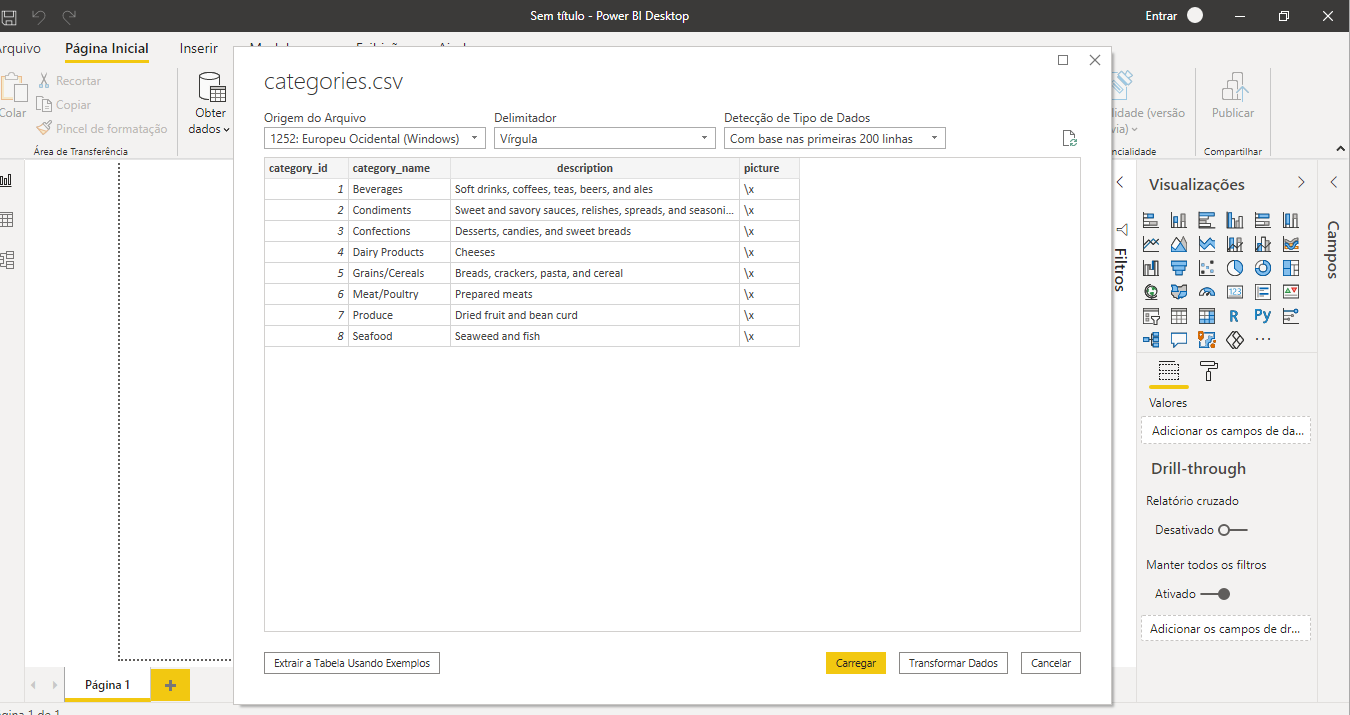
* Exportar as tabelas categories, customers, employees, order\_details, orders, products, shippers e suppliers para o formato .csv usando o pgAdmin4

###### Import/Export do pgAdmin4...





* Importar os arquivos .csv. Se for necessário, tratar os dados no Power BI



* Checar relacionamento entre as tabelas



* Construir um dashboard no Power BI



### Referências

# O que é Data Storytelling?

Data Storytelling é um conjunto de técnicas que orientam uma apresentação de informações e insights de dados para um determinado público.É a forma de contar a história dos seus dados. E essa história auxilia em manter a atenção e compreensão dos dados pelo público.

**

Na construção do Data Storytelling é importante além de contar história, saber analisar dados, dessa forma essa metodologia nasce de uma demanda da área de Data Science.

## Como criar seu próprio Storytelling

Perguntas que precisamos responder para guiar uma apresentação:

* PÚBLICO-ALVO

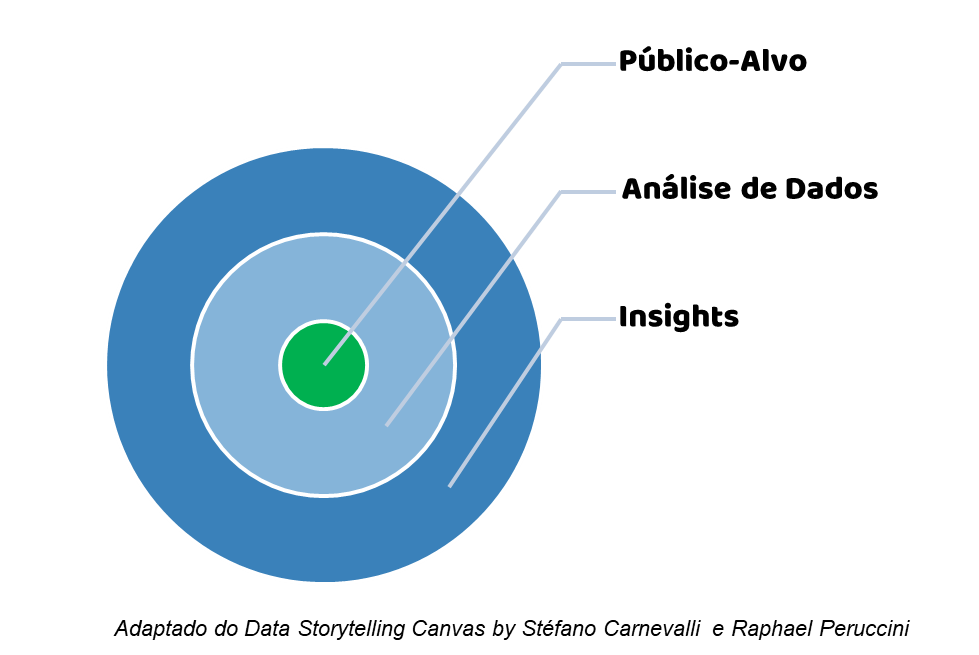
Quem têm interesse em visualizar melhor esses dados. Qual perfil. Qual cargo?

* ANÁLISE DOS DADOS

Que informações podem ser extraídas desses dados. Os dados possuem credibilidade?

* INSIGHTS

Como essas informações podem ajudar o público-alvo?

**

Essas respostas nos ajudam a construir uma história que irá manter a atenção do público-alvo.

## Desenvolvendo Storytelling

Para o desenvolvimento dessa história podemos utilizar o conceito adaptado da Jornada do Herói (Joseph Campbell, 1990):

1. Introdução: o contexto geral das informações.

2. Conflito: qual o problema e o objetivo que queremos alcançar.

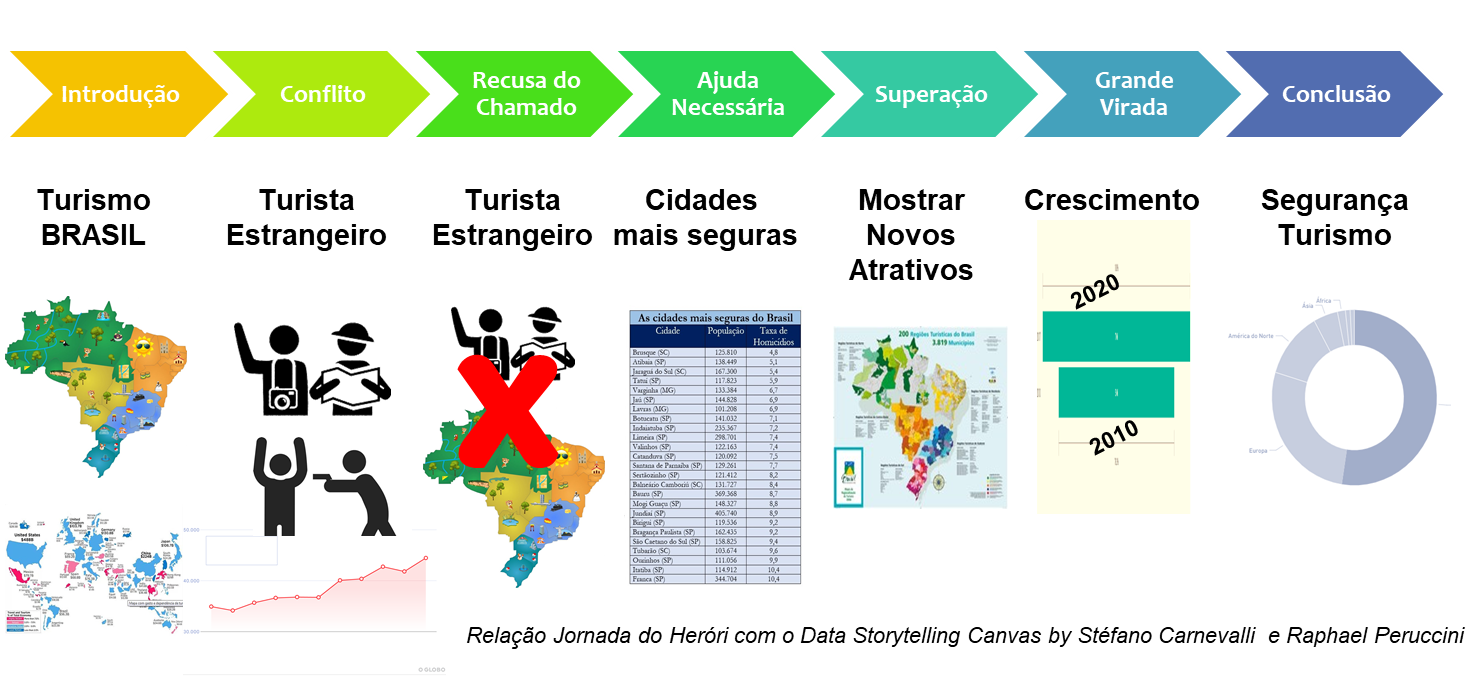
3. Recusa do chamado: o que está nos dificultando seguir em frente (falta de dados, o problema parece muito grande, a meta parece muito distante).

4. Ajuda necessária: Do que preciso. Quais dados nos ajudam a seguir em frente. Quem mais pode nos ajudar.

5. Superação: Que informações posso apresentar para poder seguirmos em frente (que outras informações facilitam atingir as metas).

6. Grande virada: Alcançando os objetivos (1º insight).

7. Conclusão da trama: o que precisa ser feito (2º insight).



O processo de desenvolvimento de uma apresentação requer ajustes da história de acordo com o público-alvo e seus objetivos. Treinar e ensaiar, apresentando para outras pessoas também ajuda. E algo que é muito importante é a credibilidade das fontes de dados, que são um dos pontos essenciais para manter a atenção do seu público.

## Storytelling conhecidos

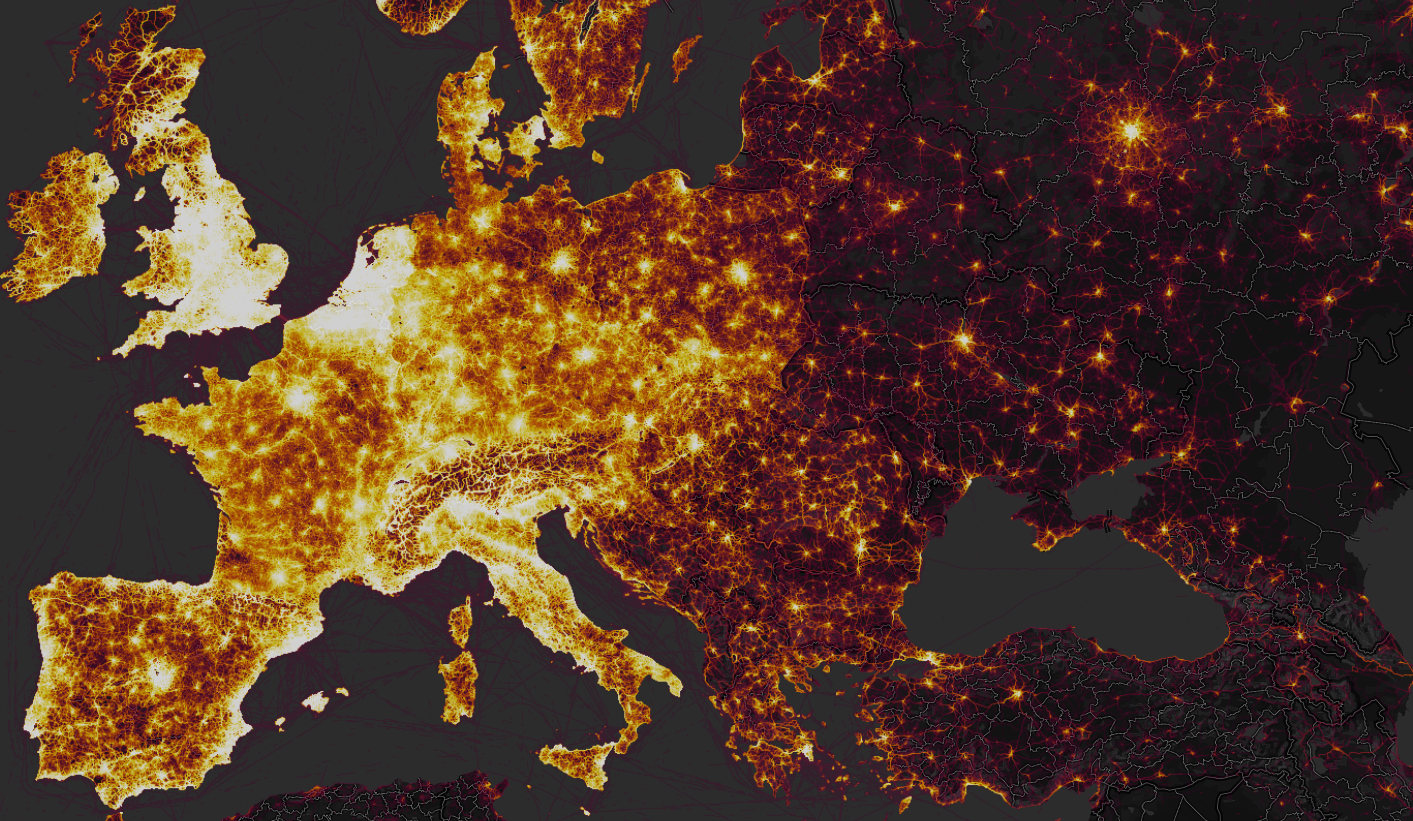
* [FiveThirtyEight.com](https://fivethirtyeight.com/)

FiveThirtyEight, às vezes representado como 538, é um site americano que se concentra em análises de pesquisas de opinião, política, economia e blogs de esportes.

**

* O heatmap da [Strava](https://www.strava.com/)

O Heatmap global mostra o "calor" gerado por atividades públicas e agregadas nos últimos dois anos. Visite strava.com/heatmap para ver o heatmap da sua área. O heatmap não é atualizado em tempo real, ele é uma fotografia de dados históricos.

**

* [R2D3](http://www.r2d3.us/uma-introducao-visual-ao-aprendizado-de-maquina-1/)

O site ensina Machine Learning utilizando elementos visuais que, inclusive, possui versão traduzida para Português. Embora seja um tutorial, a história está lá sendo contada.

**

## Ferramentas para Storytelling

* Jupyter Notebook
* RStudio
* Markdown
* Shorthand
* Datawrapper

### Informações Adicionais

[Storytelling com dados da Cole Nussbaumer](https://www.amazon.com.br/Storytelling-com-Dados-Visualização-Profissionais/dp/8550804681/ref=asc_df_8550804681/?tag=googleshopp00-20&linkCode=df0&hvadid=379805395634&hvpos=&hvnetw=g&hvrand=11870350594378386849&hvpone=&hvptwo=&hvqmt=&hvdev=c&hvdvcmdl=&hvlocint=&hvlocphy=1001750&hvtargid=pla-812777209198&psc=1)

[Blog da Cole Nussbaumer](https://www.storytellingwithdata.com/blog)