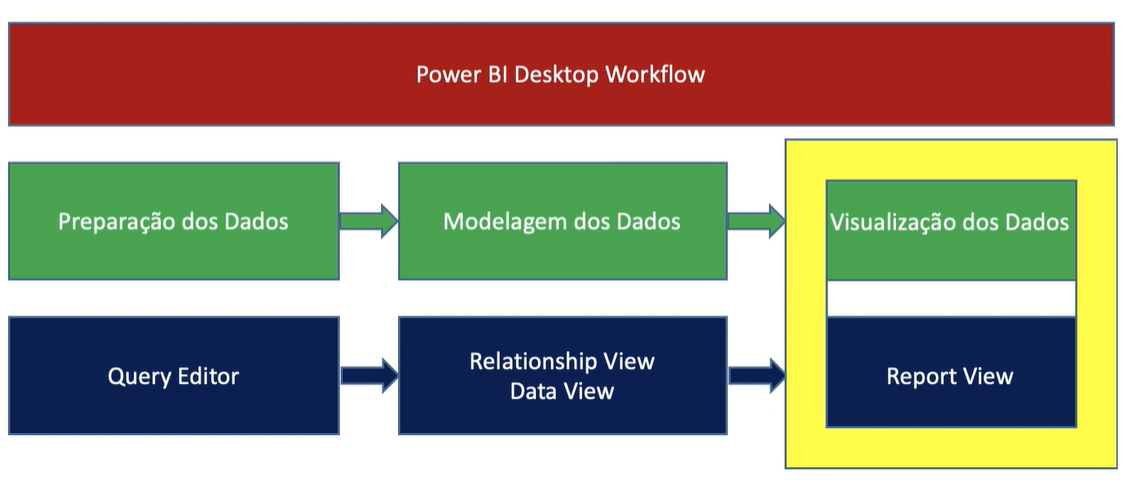
**POWER BI**

O Analista de Dados, não constrói gráficos, ele modela os dados, elabora como será feito o relacionamento, como a modelagem será implementada e como o relacionamento ocorrerá. É dedicação de tempo em modelagem, no entendimento dos dados e os relacionamentos entre eles. É não entregar as informações erradas.

# Workflow

Preparar os dados 🡪 Modelar os Dados 🡪 Visualizar os Dados



# Cardinalidade

Cardinalidade é forma que implementa o relacionamento entre as tabelas.

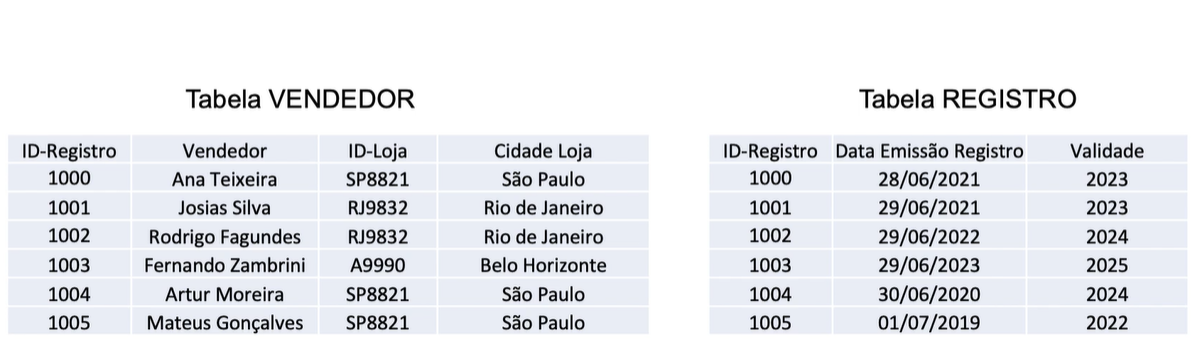
## Um para muitos (1:\*)



Cada produto é único em sua tabela de origem, porém, um produto pode ser vendido uma ou mais vezes. O produto 1000 foi vendido duas vezes. Inclusive, por vendedores diferentes.

Na tabela de PRODUTOS, o ID-Produto é a PK (Primary Key), enquanto na tabela de VENDAS, o ID-Produto é uma FK (Foreign Key).

## Um para um (1:1)



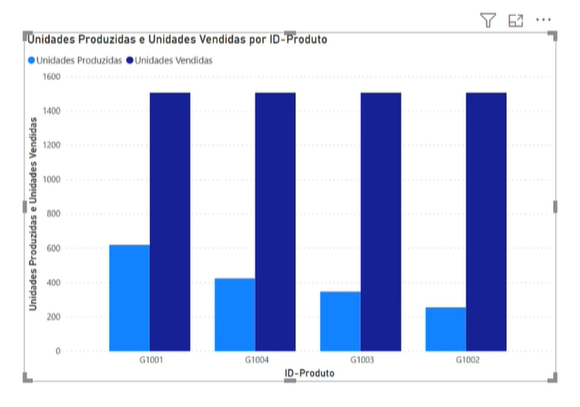
Um vendedor só aparecerá uma única vez na tabela de VENDEDOR e uma única vez na tabela de REGISTRO. Porém, podemos considerar que o vendedor entrou e saiu e ai poderia mudar o ID-Registro, adicionando apenas uma única flag fornecendo a informação de retorno ou saída do funcionário.

## Muitos para muitos (\*:\*) ou Filtros Cruzados

Isso é, muitos registros de uma tabela existirão muitas vezes na outra tabela.



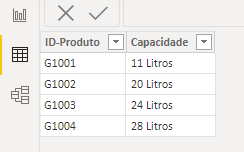
Veja que a Tabela de UNIDADES PRODUZIDAS, estou tentando relacionar com a tabela de UNIDADES VENDIDAS, o que torna que o mesmo dado existirá mais de uma e NÃO POSSUI VALORES ÚNICOS. Como por exemplo o ID-Produto G1001 que tem unidades no ano de 2014 e de 2015.



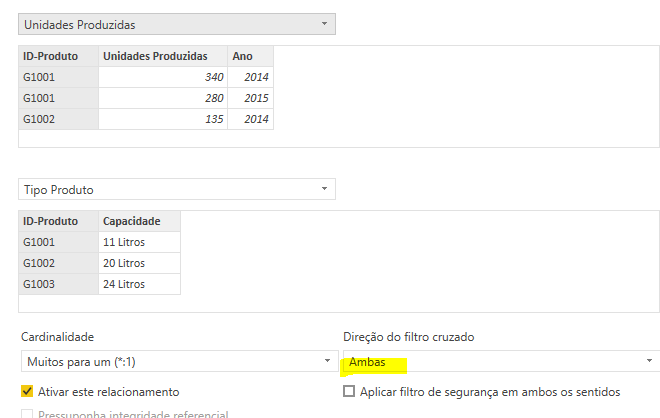
Dessa forma, apresenta o erro mostrando equivocadamente que mais unidades foram vendidas do que produzidas.

Nesse caso, utilize o “Middle Man” ou a Tabela Intermediária. É uma tabela que contém informações que existirão nas duas tabelas, mas que nela contenham valores e registros ÚNICOS.

**TABELA PRODUTOS**

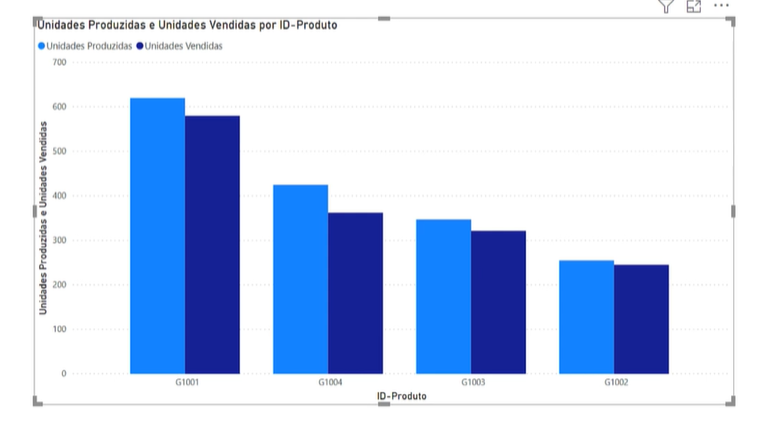


Faça o relacionamento entre tabelas que contenham registros ÚNICOS.



Lembre-se de trocar sempre o sentido do filtro para o CRUZADO.

O resultado será diferente:

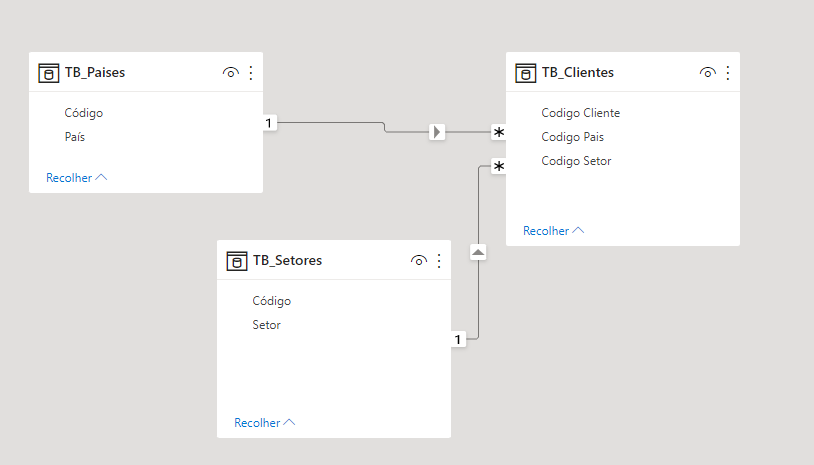


## Muitos para um (\*:1)

## Exemplos práticos

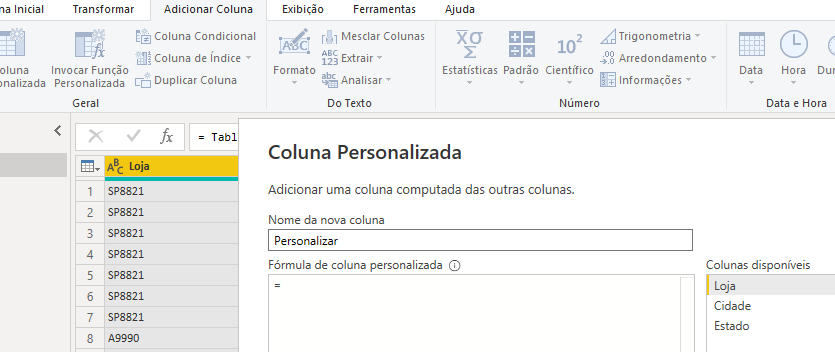
Um SETOR pode ser utilizado para vários clientes e um PAIS pode estar para vários clientes.

Embora UM CLIENTE só pode estar em UM SETOR e em apenas UM PAÍS.



# DAX – Data Analysis Expression e Power Query M (Linguagem M)

Linguagem M é utilizada para mashup de dados (Filtrar, dividir e combinar) de uma ou mais coleção de dados com distinção de maiúscula e minúscula. Para utilizar a Linguagem M, é através da TRANSFORMAÇÃO DOS DADOS. A linguagem M é mais complexa e possui mais flexibilidade do que a DAX.

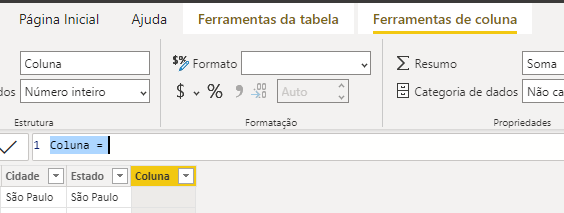


Ao tentar criar uma coluna personalizada, é necessário criar uma fórmula personalizada, com programação.

Utilizar a DOCUMENTAÇÃO OFICIAL como base.

A DAX é uma biblioteca de funções e operadores para criar e combinar fórmulas e expressões.

Para utilizar a Linguagem DAX, vá em DADOS > Adicionar Coluna (por exemplo), para utilizar a Linguagem DAX, é através de uma “barra de fórmulas” similar a barra de fórmulas do Excel.



# Gráficos

## Linha

Evolução do negócio ao longo do tempo. Linhas temporais, por exemplo estudar quantas pessoas foram admitidas ou demitidas ao longo de 6 meses.

## Dispersão

Utilizado para estudar o relacionamento entre duas variáveis.

# PostgreSQL

## Alterando o Collation do DB

Para alterar o “idioma” do collation do sql, para que não tenha problema com as datas (dates):

*ALTER TABLE mkt."TB\_VENDAS" ALTER COLUMN "DT\_COMPLETA" SET DATA TYPE CHARACTER VARYING (20) COLLATE "pt-BR-x-icu"*

O COLLATE serve para alterar para a “localidade

# Estatística

O que é Estatística?

É um conjunto de métodos para se analisar dados, com ferramentas que proporcionam possiblidades de se tomar decisões em um ambiente de incertezas.

Pode ser considerara a Ciência de Aprendizagem a partir de dados.

Dados está no centro da 4 revolução industrial

Os dados em si,são matéria prima bruta.

Técnica de análise de dados que são fornecidas pela estatística. Que gera dados para gerar informação.

## Probabilidade

Estudo da aleatoriedade e incerteza

## População e Amostra

A população é o “total de indivíduos distintos”

A amostra é uma “parte da população”.

No início do estudo, geralmente se define a população e se retira uma amostra para ser examinada.

Com base nas amostras, são gerados estudos sobre uma população para gerar o conhecimento.

Cuidado ao coletar amostras para não acontecer um estudo tendencioso. Por exemplo, estudar algumas características de pessoas no Japão, no Brasil, Estados Unidos.

Medir uma população inteira é muito custoso, leva-se tempo e torna-se inviável.

## Censo

É a técnica que seleciona e avalia todos os elementos da população quando se realiza uma pesquisa.

## Estatística Inferencial

Através das análises das amostras, conseguimos gerar INFERÊNCIAS.

É possível fazer uma avaliação muito precisa sobre uma população inteira. É possível fazer inferências sobre uma população.

Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente

## Tamanho da Amostra

Cuidado com o tamanho da amostra, para que a amostra reflita o tamanho da população. Exemplo da colher de sopa para se medir o tempero e o sal, mas que a sopa está em um caldeirão imenso.

## Dificuldades na Amostra

Todos os elementos precisam ser sorteados, números suficientes, sem repetições, (coletar, tamanho da amostra e randomização da amostra)

## Amostragem Probabilística

Métodos e ferramentas mais objetivos, sem influência da pessoa que está coletando a amostra.

Os elementos são apresentados de forma aleatória.

## Amostragem Não Probabilística

Existe uma influência direta de quem está realizando a coleta das amostras.

É quase que por “Conveniência”.

## Parâmetros x estatísticas

O parâmetro é uma característica sobre a população. São valores calculados usando dados da população. A média, o desvio padrão da população.

A estatística é o resultado de uma amostra. “A estatística de desemprego do Brasil foi de x%”. É o resultado da análise da amostra. Isso é estatística.

A amostra é usada para calcular a estatística, que usa para estimar os parâmetros dela, e ai se compreende uma população.

Diagrama

Descrição gerada automaticamente

## Dados Primários x dados secundários

Os dados primários são analisados por quem faz a coleta, são mais confiáveis. Porém, custa muito mais e demanda uma equipe maior.

Os dados secundários são coletados por terceiros e podem não ser tão confiáveis. Nesse caso, ele apenas é mais vantajoso apenas pela questão financeira.

## Observação

É uma ocorrência de um item de dados, um registro, uma linha em uma tabela, um evento, uma instância.

## Variável

Podem ter valores numéricos ou não, uma coluna em uma tabela, atributos.

Diagrama

Descrição gerada automaticamente

### Tipos de Variáveis

A coluna Peso, é representada por valores numéricos. É possível realizar a média dos pesos dos indivíduos.

#### Qualitativas

Termos descritivos (cor dos olhos, tipo sanguíneos, cor do cabelo, pele)

**Nominais**: Sem ordem natural (Sexo da pessoa) não existe hierarquia

**Ordinais**: Uma ordem natural (índice da aprovação do político, ensino médio, classe social da pessoa, uma ordem da fila)

#### Quantitativas

Representados por valores numéricos: peso, número de filhos

**Discreta**: Valores podem ser contados, não podem existir valores decimais (números de filhos)

**Contínua**: Podem ser medidos (altura (1,5m), peso(68.8 kg), salário pessoa (R$ 4.533,99)

A idade: 5 anos, 18 anos, 3 anos, 60 anos é **Quantitativa Discreta.**

A idade 0 anos, 1 ano, 3 anos, 15 anos pode ser de faixa etária, logo **Qualitativa Ordinal.**

## Medidas de Posição

São valores que evidenciam a tendência de concentração de dados. As mais “famosas”. Elas são calculadas no conjunto de dados.

### Média

É a soma de todos os registros das variáveis dividido pelo total de registro

A média é impactada por valores extremos (outliers)

### Mediana

É o conjunto de dados que divide em duas partes com a mesma quantidade de dados:

Se o conjunto de dados for par: (Mediana = (N + 1)/ 2

Se o conjunto de dados for ímpar: (Mediana = (N / 2) + 1

### Moda

É o valor que mais se repete, o de maior frequência no conjunto de dados.

Gráfico, Diagrama

Descrição gerada automaticamente

Ao estudar as Medidas de Posição, é possível realizar estudos e suposições sobre uma variável.

## Medidas de Dispersão

Elas descrevem a amplitude dos dados, ou seja, o quanto eles estão espalhados ao longo do conjunto de dados. São eles: Variância e Desvio Padrão

### Variância

Mede a amplitude, mostra o quanto os valores estão distantes da Média.

### Desvio-padrão

É usado para medir a variabilidade dos dados em um conjunto. É a distância média da média.

Julio:



Fernanda:

Tabela

Descrição gerada automaticamente

Interface gráfica do usuário, Aplicativo

Descrição gerada automaticamente

Pode-se entender o Desvio Padrão da seguinte forma:

Imagine que uma concessionária vende carros.

Gráfico

Descrição gerada automaticamente

O valor do carro com maior valor foi de R$ 181.25Mil e o mais barato custou R$22.500 mil.

O desvio padrão de 45.220,07 sugere que na verdade, que a maioria dos carros podem custar R$ 45.220,07 reais para cima ou para baixo da **Média**.

## Coeficiente de Variação

Mede o desvio padrão em termo de % da média.

Quanto maior o CV, menor consistência no conjunto de dados.

Serve para comparar os dois conjuntos de dados

CV = ( S / X) \* 100

## Medidas de Posição Relativas

Compara a posição de um valor com outro valor em um conjunto de dados.

### Percentil

Uma escala de 0 a 100.

Busca de uma posição e indica qual é. Como abaixo, buscar o Lucas em uma escala de altura, significa que 80% dos casos estão aqui ou abaixo dos80%

Desenho de personagem

Descrição gerada automaticamente com confiança média

#### Percentil x Porcentagem

Não é a mesma coisa, são conceitos diferentes, a Porcentagem e um por cento de um todo e o percentil é a posição que ocupa.

### Quartil

Fundamental para a análise de um Boxplot.

Gráfico, Gráfico de caixa estreita

Descrição gerada automaticamente

Dividem uma tabela de dados em 4 partes iguais.

Texto

Descrição gerada automaticamente

O segundo quartil é exatamente igual a MEDIANA.

Gráfico, Diagrama, Gráfico de caixa estreita

Descrição gerada automaticamente

O conjunto de dados está mais próximo do Valor Mínimo ou do Valor Máximo? A média está distante da Mediana?

# Linguagem R com Power BI para Análise Estatistica

Baixe a linguagem R, depois o R Studio e depois o R Tools(essa última precisa ficar gravada na máquina dentro da raiz C).

Interface gráfica do usuário, Texto

Descrição gerada automaticamente

Depois de criar o script lá no R studio, precisa colocar o arquivo que será trabalhado no mesmo diretório onde está o script do R.

Tabela

Descrição gerada automaticamente

Códigos:

#**Define a pasta de trabalha onde o script está** alocadosetwd("D:/workspace/1\_DSA\_POWER\_BI/cap12R")

#**Evidencia o local do script salvo**  
getwd()

# **Carrega um dataset para trabalhar**  
vendas <- read.csv("Vendas.csv", fileEncoding = "windows-1252")

# **Resume o Dataset**  
View(vendas) #abre uma grid no R Studio com os dados

#**Mostra as variáveis e seus tipos de dados**  
str(vendas) #as vezes DATAS podem vir como “CHARACTER” precisa arrumar.

Texto

Descrição gerada automaticamente

#**Mostra o sumário de alguma coluna em específico (como media,moda,mediana)**  
summary(vendas$Valor)

Texto

Descrição gerada automaticamente

summary(Vendas$Custo)

Texto

Descrição gerada automaticamente

## Média Aritmética com R Studio

#**Calcula a Média**  
?mean #pergunta pro R Studio e ele fornece uma ajuda

#**Calcula a média de uma coluna do dataset**mean(vendas$Valor)

mean(vendas$custo)

## Média Ponderada

?weighted.mean

weighted.mean(vendas$Valor, w = vendas$Custo)

## Mediana

#Mediana

median(vendas$Valor)

median(vendas$Custo)

Interface gráfica do usuário, Texto

Descrição gerada automaticamente

## Criando uma função no R Studio (a Moda)

moda <- function(v){

valor\_unico <- unique(v)

valor\_unico[which.max(tabulate(match(v, valor\_unico)))]

}

#Obter a moda:

resultado <- moda(vendas$Valor)

print(resultado)

Interface gráfica do usuário, Aplicativo

Descrição gerada automaticamente

resultado\_custo\_moda <- moda(vendas$Custo)  
print(resultado\_custo\_moda)

Interface gráfica do usuário, Aplicativo

Descrição gerada automaticamente

## Criando um gráfico no R Studio (GGPLOT)

#Instala um pacote  
install.packages("ggplot2") #única vez para instalação  
library(ggplot2) #**carrega o pacote na sessão do script (sempre que abrir o script)**

Texto, Carta

Descrição gerada automaticamente

ggplot(vendas) +  
 stat\_summary(aes(x = Estado,  
 y = Valor),

fun = mean,

geo = "bar",

fill = "lightblue",

col = "grey50") +

labs(title = "Média de Valor por Estado")

Gráfico, Gráfico de mapa de árvore

Descrição gerada automaticamente

## Medidas de Dispersão no R Studio

#ESTATÍSTICA BÁSICA

#PARTE 2 - Medidas de Dispersão

#Seta o local do script

setwd("D:/workspace/1\_DSA\_POWER\_BI/cap12R")

getwd()

#Carrega o dataset

vendas <- read.csv("Vendas.csv", fileEncoding = "windows-1252")

# Resumo do dataset

View(vendas)

str(vendas)

#Resumo das estatísticas básicas

summary(vendas)

#Variância

var(vendas$Valor)

#Desvio-padrão

sd(vendas$Valor )

## Medidas de Posições Relativas

Faz uma varredura das estatísticas básicas de duas colunas ou mais.

summary(vendas[c('Valor', 'Custo')])

Tela de celular com texto preto sobre fundo branco

Descrição gerada automaticamente

#Estatística Básica

# Parte 3 - Medidas de Posição Relativa

#Definie a área de trabalho do script R

setwd("D:/workspace/1\_DSA\_POWER\_BI/cap12R")

getwd()

vendas <- read.csv("Vendas.csv", fileEncoding = "windows-1252")

#Resumo dos dados

head(vendas)

tail(vendas)

summary(vendas$Valor)

summary(vendas[c('Valor', 'Custo')])

#Variáveis numéricas

mean(vendas$Valor)

median(vendas$Valor)

#**Informa os quartis**

quantile(vendas$Valor)

#**Calcular o Percentis: percentil de1% e 99%**

quantile(vendas$Valor, probs = c(0.01, 0.99))

#**Define um limite**

quantile(vendas$Valor, seq(from = 0, to = 1, by = 0.20))

#**Diferença entre Q3 e Q1**

IQR(vendas$Valor)

#**Do valor mínimo ao valor máximo do atributo**

range(vendas$Valor)

#**Confere se está o menor e maior mesmo**

summary(vendas$Valor)

#**Calcula a diferença entre o maior e o menor**

diff(range(vendas$Valor))