

Power Bi

Fazer a pergunta certa para os dados

UN **encod**: quando comegor o arquivo, lembrar de ajustar o encode para corrigir acentuações.

Detectar do **tipo** de Dados:

Baseado no conjunto interno

Antes de comegor os dados você tem a opção de **editar** os dados.

Se der 2 s
Clique pode alterar o título
do topo da coluna

	Nome	L1	L2	L3	L4
1					
2					
3					
4					
5					
6					

Você pode editar o tipo de dado
calendário
texto
número
etc

Fechar e aplicar

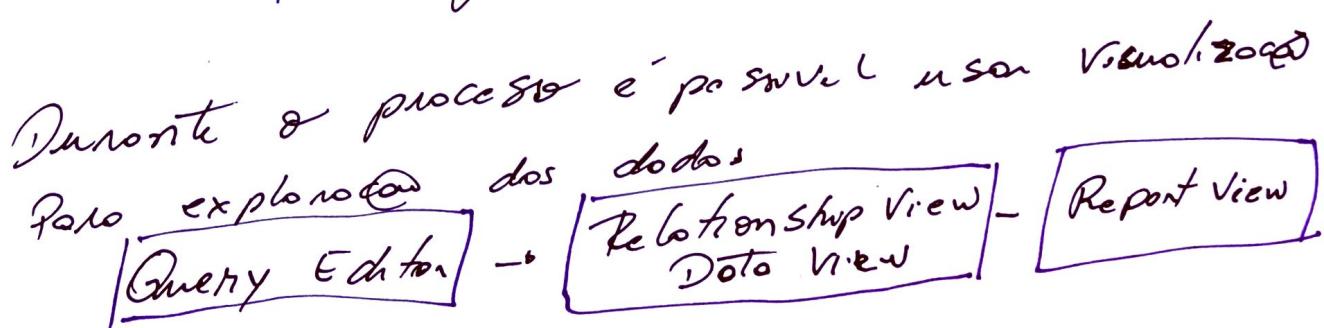
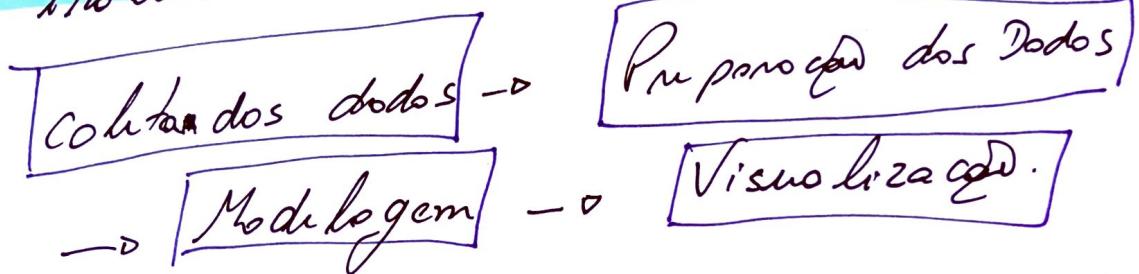
— é possível adicionar barra **média** no Gráfico

— Pode se alterar a **visualização** para várias tipos de Gráficos, inclusive **mapa** com países e estados.

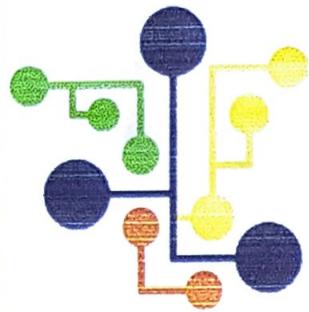
— é possível descer a hierarquia = tipos:
Ano → trimestre → mês → dias

- temos a opção de ver especificamente o ano de 2016
1º tri 2º tri ...
ou algo genérico como:
1º trimestre de todos os anos
- menu = dados / Form...
 → mostrar nível seguinte
 → Exibir nível seguinte

A definição do problema é analisar comportamento e trânsito.



trade off = escolha.



Data Science Academy

www.datascienceacademy.com.br

Microsoft Power BI Para Data Science

Estudo de caso
Definição do Problema de Negócio

Área técnica + Área de negócios
é a melhor opção

Você é Analista de Dados na empresa XYZ Corporation International, uma revendedora de automóveis de luxo com sede em São Paulo. A empresa começou sua operação no Brasil em 2012 e atua nos quatro estados da região sudeste mais os estados do Paraná e Bahia.

Seu gerente vai apresentar os resultados da equipe comercial para o novo CEO da empresa e precisa da sua ajuda para construir um Dashboard que represente os dados de vendas no período de 2012 a 2015.

Sua fonte de dados é um arquivo Excel com dados coletados do sistema de vendas e CRM da empresa, com as seguintes colunas:

Coluna	Descrição
DataNotaFiscal	Data de emissão da nota fiscal
Fabricante	Fabricante do veículo
Estado	Estado onde foi realizada a venda
PrecoVenda	Preço de venda do veículo
PrecoCusto	Preço de custo do veículo para a empresa
TotalDesconto	Total de Desconto fornecido sobre o preço de venda
CustoEntrega	Custo de entrega do veículo ao proprietário
CustoMaoDeobra	Custo de Mão de Obra (secretária, mecânico, etc...)
NomeCliente	Nome do cliente que comprou o veículo
Modelo	Modelo do veículo
Cor	Cor do veículo
Ano	Ano de fabricação do veículo

Seu gerente precisa das seguintes informações:

- 1- Total de Vendas por Ano
- 2- Custo de Entrega do Veículo Por Fabricante
- 3- Custo de Mão de Obra Por Estado
- 4- Total de Vendas Geral e Matriz de Vendas

Além disso, pode ser interessante, se o CEO puder visualizar o total de vendas por estado e se as vendas estão acima ou abaixo da média. Seu gerente já sabe que um assunto será abordado pelo CEO durante a apresentação. O CEO está avaliando se continua ou não com a venda de automóveis da marca Jaguar e ele gostaria de saber como evoluíram as vendas de automóveis deste fabricante por ano e por estado.

Seu trabalho é fazer isso acontecer!

*trabalho de Bi =
olhar para o passado
para entender o
Problema*

Análise preventiva = olhar para o futuro

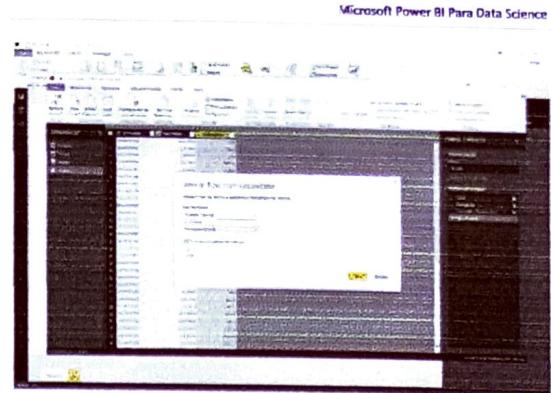


Data Science Academy

www.datascienceacademy.com.br

Microsoft Power BI Para Data Science

Localização e Números Decimais

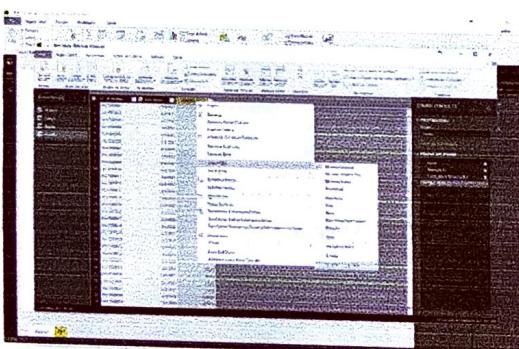


Selecione Número Decimal e então escolha Português Brasil. Clique em Ok.

Microsoft Power BI Para Data Science

Dependendo do formato dos dados de entrada, o Power BI pode não reconhecer os números decimais.

Isso não é um problema ou erro, apenas precisamos informar ao Power BI qual formato de número decimal deve ser utilizado, o que é feito através da configuração de localização.



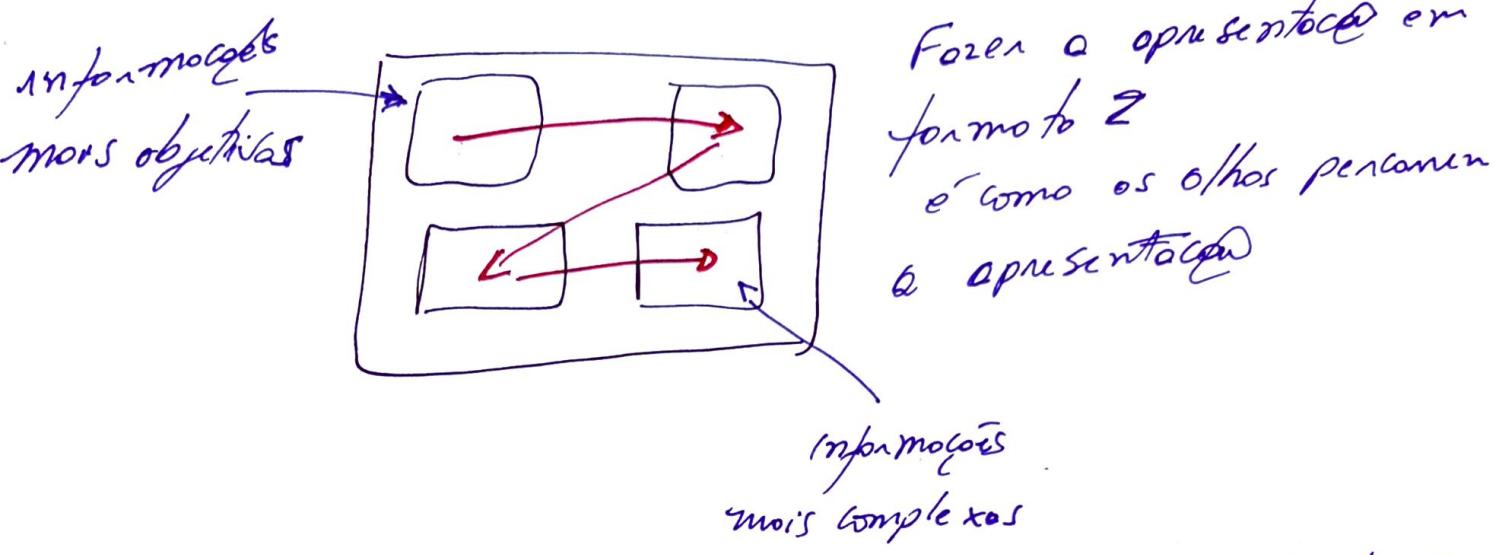
No Power BI, trabalhando com o Query Editor, clique com o botão direito na coluna que contém números decimais e clique em Alterar Tipo e então clique em Usando a Localidade, conforme tela acima.

Dashboard

Story telling = contar história

Se você precisar interpretar o gráfico
Significa que ele está muito complexo
Simplificado é a maior sofisticação

Tabela → Formatação → Estilo → Cabeçalho em negrito



Web Scraping = coletar informação diretamente das páginas web

Preciso sem servidor, sem vincula

menu → Arquivo → opções e dif. → opções
→ Definição Regional

Obtenha dados → Web → colar link da tabela

- Depois de conectado
- Vista da tabela / Vista web
 - ↳ Seleciona a tabela que deseja importar

É preciso organizar os dados dos tabelos que podem vir com essas linhas desorganizadas, como:

titular das colunas na 1ª linha.

→ Você tem a opção: utilizar a 1ª linha como cabeçalho

→ Você pode remover colunas vazias ou com Null

A direita você tem um exemplo de atividades feitas e você pode excluir essas atividades

Soltando = Rollback

Fechar e aplicar

Pode-se converter várias colunas em linhas menu: transformar - com a Dinamização de coluna

1	2	3	4	5

→

1	
2	
3	
4	
5	

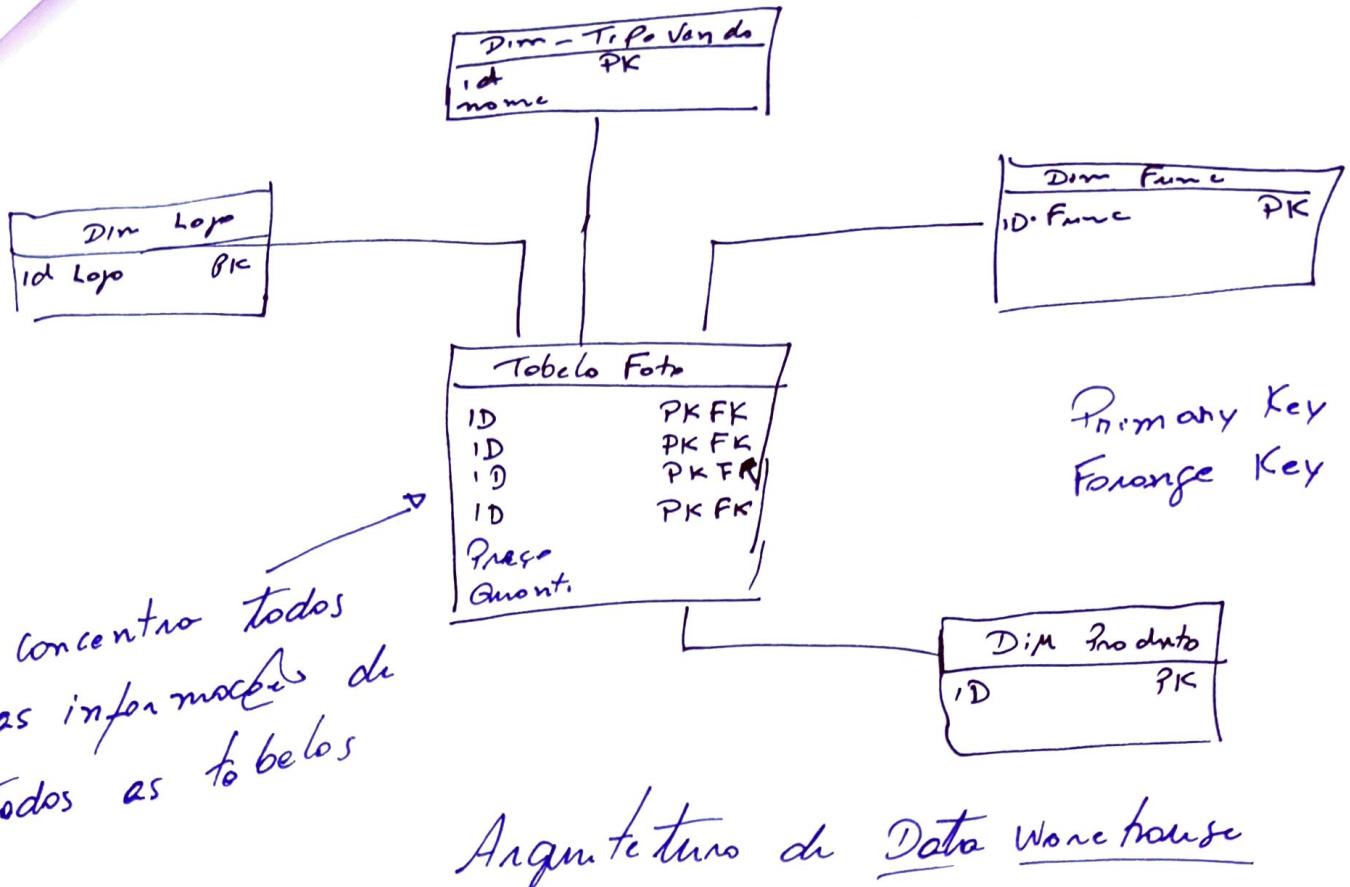
menu =

transformar → detectar tipo de dados

identifica o tipo de dado de cada

Coluna

Star Schema



Ventabelo

Design of data base

Criar Nova tabela a partir de uma Principal

Carta direta para no principal

→ Editor consulta

No editor → corts esquerdo

→ dois critérios no nome do tabela P/
alterar o nome.

→ botão direito P/ duplicar a tabela

→ botão direito na area para **criar**
um **grupo** de tabelas

→ ícone à esquerda no topo
↳ escolher colunas

Duplicando tabelas

Editor consultas → Escolher colunas

Sempre duplicar tabela principal

Tabelas Dimensionais

- Dim - Produto
- Dim - Região
- Dim - Vendedor
- Dim - Tempo

star schema

tabela dimensionada
não pode ter itens
duplicados

Tabelas Fato

- Fato - Vendas
 - ↳ ID - Produto
 - ID - Loja
 - ID - Vendedor
 - ~~ID - Data Venda~~
 - Valor Venda

BI = Entender o que aconteceu no passado

Gerenciar Relações

Você pode receber um aviso de que a tabela possui itens duplicados

Clicar em: Base =



→ detectar automaticamente

→ Editor Relações:

Condição lida da 1 Paro 1
1 Paro * (muitos)
* Paro *

Novo Relacionamento

- Escolher tabela dimensional - ID

Escolher tabela Fato - ID

Adicionar nova medida

à dimensão

- clicar em opções da tabela (...)

Foto Venda
ID
ID
ID
ID
Total Venda

→ nova medida

→ nome da coluna

Nova coluna de valor
Novo nome

→ Sum (coluna determinada)

Limpeza e transformação de dados

É preciso verificar as configurações regionais
se o Power BI está configurado de forma
compatível com o ambiente

No menu =D

→ Arquivo

→ opções e definições

100,00

→ opções

→ Definições Regionais

→ inglês (USA)

Resultado: os **Volumes** estão com as
coisas da maneira correta

Ajustar Unicode:
(UTF-8)

Acentos

"São Paulo"

Corrigir Vários organismos de uma vez

- Obter dados
- Posta
- Passar comando do posta
- Combinar e editar

Se quiser adicionar algum organismo.
é só jogar o nome do organismo dentro da
posta e clicar em atualizar

textos com caracteres não formados como tab
menu → transformar → Formato →
contor (tum)

Remove espaço's no fim e inicio
→ limpar (clean)

tira os tabs do inicio do texto

Você pode substituir por exemplo ---

3 espaço's por 1 espaço
menu → transformar → Substituições

Alterar o formato das palavras, por exemplo

Se x = mostrando e f = numero

Você pode selecionar Formato e colocar

cada palavra em ~~mostrar~~ mostrando

Moscarino Fornmann

Metadados

no canto direito

Etapas aplicadas = onde as suas fotos ficam listadas.

Você pode clicar no ícone de engrenagem e adicionar uma observação explicando qual foi a alteração feita.

P/ você lembrar em quando entre pessoas pessoas, saber o que foi feito

- Definir regras de negócio
- coletar dados
- Limpeza e organização dos dados

coluna do tipo numérico que não se for operação pode ser convertida em texto

Sempre lembrar de ajustar o ~~modo~~ modo ~~tabuleiro~~ P/ verificar a separação das colunas

Split de Coluna

em Editor do Power Query

→ botar diretamente na coluna

↳ Dividir coluna

↳ Pelo Delimitador ⇒ ; , / tab etc

↳ Pelo número de caracteres

Reconhecimento de tabelas

Você pode anotar e soltar o cursor para detectar relacionamentos

converter Data configuração Regional

Se dentro da data pode estar em um formato regional diferente.

Mudar tipo → utilizar Regras

→ Dado → Português Brasil



Valores com R\$

\$	1	v
----	---	---

↓
→ Utilizar Regras

→ Tipo dados = nº decimal

→ Regras = Português Brasil

Remover linhas com erro

\$	1	v
----	---	---

↓
→ Botão direto

→ Remover linha com erro

Mudar texto R\$ para número
é possível modificar R\$ para nada

também é possível alterar not agreed

Para

\emptyset (zero)

$\frac{1}{62}$ Substituir Volumes

Preencher Volumes

Categoria
Acomodado

→

Categoria
Natu
Null
Null
Null
Null

→

Categoria
Acomodado
Acomod.
Acomod.
Acomod.
Acomod.

Botar Direto
Substituir

→

Branco
Pra
Null

→

Preenchimento
Pra Branco

no mundo

↳ Adicionar Colunas

↳ Coluna Personalizada

$$IMC = \text{Peso} / \text{Altura}^2$$

$$IMC = [\text{Peso}] / (([\text{Altura}] / 100) * ([\text{Altura}] / 100))$$

Hadoop

Hadoop = HDFS + MapReduce

HDFS = armazenamento distribuído

MapReduce = computação distribuída

Baixo custo - Escalável - Tolerante a falhas -

MapReduce = Flexibilidade = processa dados de forma independente, estruturados e não estruturados

comprobidade = processamento jobs em paralelo, caso um falle outros não são afetados

Acessibilidade - Suporta Java, C++
Python, Apache Pig

HDFS → (Hadoop Distributed File System)

- Gerenciamento de Várias máquinas - cluster
- tolerante a falhas
- conjunto de computadores para armazenamento grandes arquivos
- WORM - (Write once, Read many times)
- Leitura do arquivo inteiro e não apenas o 1º registro

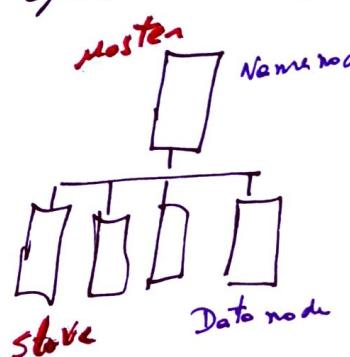
HDFS possui 2 tipos de NODES

~~Master~~ ou namenode

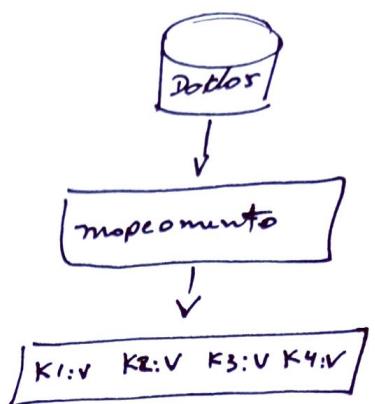
(master node)

Data node

(worker node)

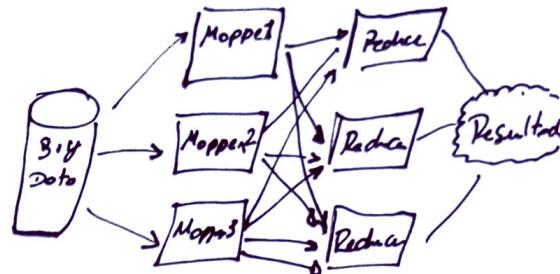


Mapreduce = modelo de programação p/ processar grandes volumes de dados



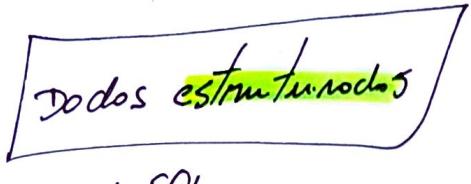
converte dados em pares de chave - valor

K = Key
v = value

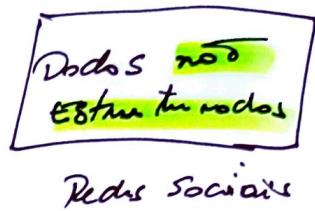
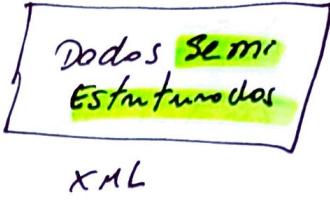


Seek time x transfer Rate

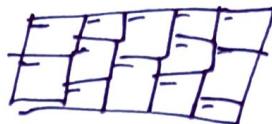
RDBMS



Map reduce
Hadoop



tree map



Atendendo com os relocaamentos, cosa seja feito
inconscientemente o gráfico não do certo.

Séries temporais



mostrar próximo nível

→ Ano → trimestre → mês → Día

É possível alterar os valores apresentados no gráfico com **soma**, **média**, **Valor máximo**

Exemplo: Gráficos desempregados



só precisa **anotar sexo** e **legenda**

Filtros = Slice

você pode aplicar filtros no gráfico como

Período Ano

selecionar apenas **Masculino** ou apenas **feminino**

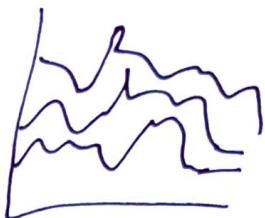
Slice = podemos modificar o gráfico mas se quisermos dor isso opção ao usuário escolhermos a opção **slice**.

Gráfico de Ano



Pois a visualização
fazcer muito confuso
quando tem muitos
dados

slice



Sexo
 Feminino
 Masculino



Selecionar o
 Botar "Feminino"

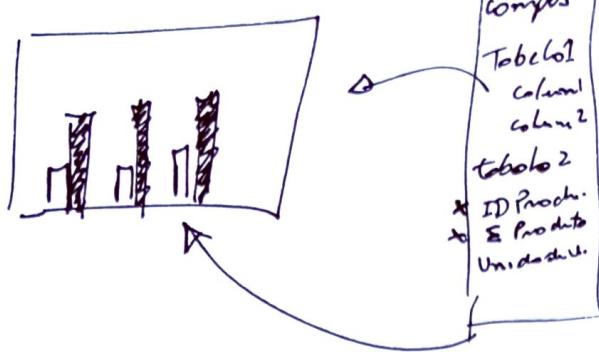
Segmento com de
 Dados

Idade	
<input type="checkbox"/>	15 - 22
<input checked="" type="checkbox"/>	23 - 30
<input type="checkbox"/>	31 - 50
<input type="checkbox"/>	51 - 60

Período 01/01 2010

Cross Filter

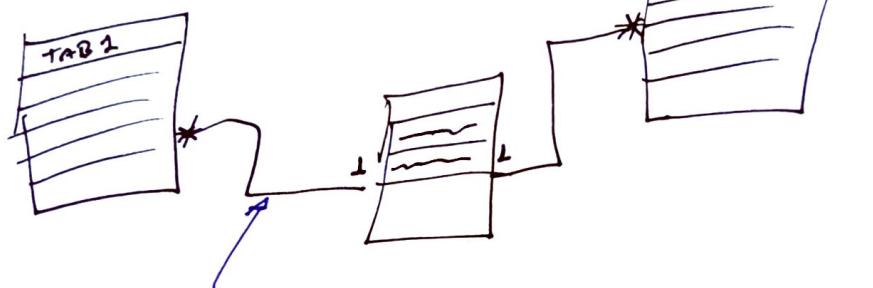
importante!



Você pode amostrar valores
 de duas tabelas distintas

No campo valor, onde você pode soltar os valores, você pode determinar se o valor fico abaixo ou acima, isto modo o gráfico.

Relacionamento muitos para muitos



- 2 Clique
 - direção do filtro cruzado
- ↳ ambos

! Erro de relacionamento muitos para gerar gráficos
 Complexo formando erros

Gênesis Relacionamentos

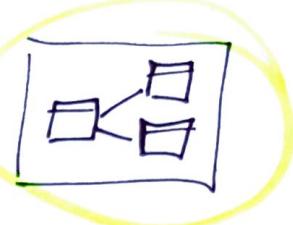
Editor de relacionamento

Vendas

ID-Produtos

Data Venda

Valor Venda



Produto

ID-Produto

Produto

condições de lidação

muitas para L (*-1)

chaves de filtro com 2000

único

muito importante a condição de estan
correta

Modelagem DAX x M-language

Data View (esquema)

DAX = Fórmulas Excel = Nova Coluna → Banco de formula

M-language = Query editor = coluna personalizada

DAX - Data Analytics Expression

Criar nova coluna com itens. - Query Editor
m-language

- Coluna personalizada

L0 = [Produto] & [serial number]

↑ E comercial (concatenar)

DAX = Data analysis Expression

Relatório

Data view

Relacionamentos

Quando digita "[" aparece automaticamente as opções de colunas

Proj 5

Coluna = []

Você pode digitar (concatenate)

concatenate ([Produto]; [Série Número])

Coluna
Coluna lida

É semelhante ao excel, Formulas para cálculos

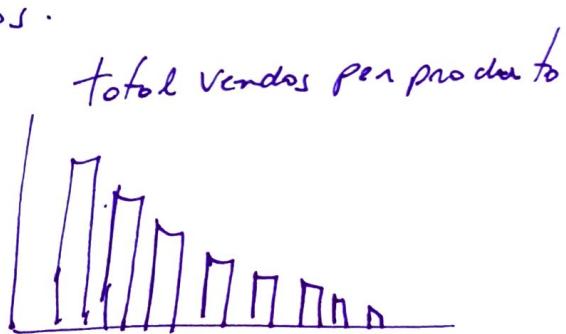
Montagem = divide([Preço Custo]; [Valor vendido]; 10)

Coluna = IF([Montagem] > 0,5; "Positivo"; "Falso")
Label

Nova medida

Total Vendas = sum(custos [Valor vendido])
não cria coluna, apenas soma

→ é possível fazer a combinação dessas medidas com outras colunas.

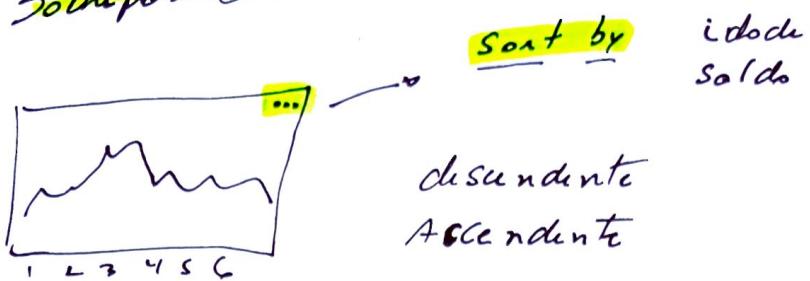


OBS: anotar coluna para eixo

Formatos de Gráficos

Format = Align / Distribute

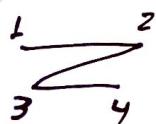
Sobreposição = botões → variação ou recuar



Você pode mudar o fundo do Gráfico

Dashboard tem origem nos países de cores

- considere seu público alvo
- Conte uma História - Story Telling
- Use o topo interno
- informações mais relevantes.
- O mais importante no topo Superior esquerdo



Podemos baixar outros ~~estilos~~ estilos de Gráficos em: app.powerbi.com/vizmkt

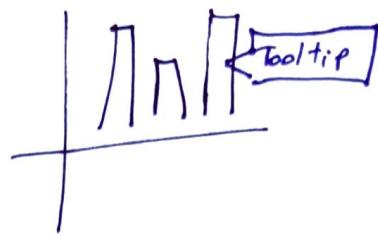
Editor interações

↳ Quando você clica em um gráfico automaticamente você afeta os outros

menu = Format → editar interações ☐ ☐

Filtros → ☐ & Nenhum

Adicionando tooltips



Visualizations

:

ToolTips

Add data field here

Você pode anotar
colunas para o área **tooltip**

Você pode alterar o que vai
ser exibido como, por exemplo
com **tags** em vez do nome dos
códigos.

Você pode usar o **soturno** de cor pra indicar
alguma informação além do tomada das cores

Slices



códigos
12
12
12
12
12
12
12
12

Você pode alterar
para **Horizontal**

OBS: desmarcando a **seleção** nesse você pode
selecionar vários códigos ao mesmo tempo

Power Bi com **Ontrack**

→ Ponto 1521 importante

givaldo.juanascor@outlook.com

Senha: Descontos 1981 ?

Márcio System
Santo System

Oracle e Power BI

Esquema = um conjunto de objetos dentro do Banco de dados

outros usuários = cada usuário é um esquema
Lo botar diretamente → Cria usuário.

Data Warehouse - star Schema

Operações ETL \Rightarrow Extração | Transformação | Limpeza

Limpeza e organização dos dados
Para inserir no Oracle

- trata os dados no Power BI
- copia a planilha para um arquivo CSV
- importa no Oracle.

Você pode fazer perguntas ao Power BI e ele responde com gráficos.

total de vendas por Estado

tem opção de insight:

faz Gráficos automaticamente.

Refresh com seu data set

Lo atualizações do data set

Dado Gateway para o refresh

Podemos agendar atualizações no Gateway

Podemos compor filhos os organizados

App Power BI

- Amostrar de como montar os Gráficos

Machine learning

tipos de aprendizado

- Supervisionada
- não Supervisionada
- Semi Supervisionada
- Por esforço
- Deep learning

Definir problema de negócio

Em busca do melhor modelo para predição.

Cada modelo para cada problema

Representação = conjunto de modelos

Avaliação = melhor modelo

Optimização = melhorar o modelo

Nenhum algoritmo é 100% preciso todo tempo

Cost function = nível de erro

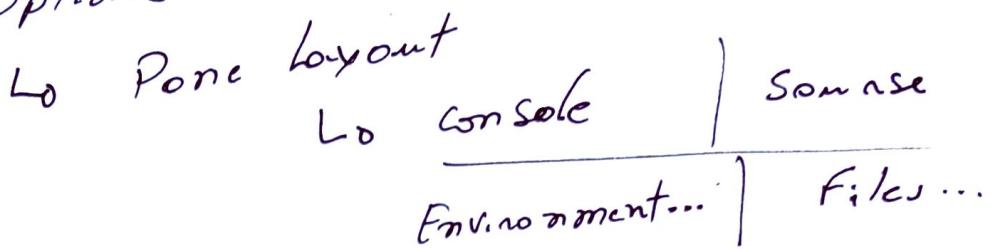
R similar a SAS

Microsoft R Server

Habilidades e script do R no power BI

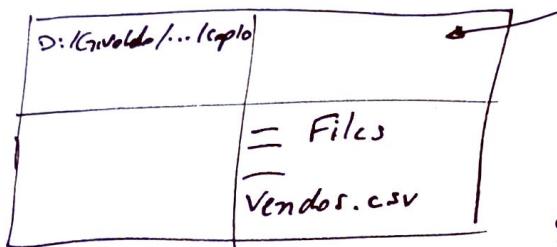
Tools → Global options

↳ Options



- No R tudo é objeto

- Session → Set Working directory → choose directory



dataset <- read.csv("vendas.csv")

↳ click em run.

plot(dataset\$valor)

Selecionar a linha execute
openas a linha

install.packages ("ggplot2") = instala pacotes

getwd() = verifica diretório

setwd("diretório")

install.packages ("pacote")

```

# Definindo a pasta de trabalho
getwd()
setwd("D:/DSA/PowerBI-DataScience/Cap10")

##### Pacotes do R #####
# Instalando os pacotes para o projeto (os pacotes precisam ser instalados apenas uma vez)
install.packages("Amelia")
install.packages("caret")
install.packages("ggplot2")
install.packages("dplyr")
install.packages("reshape")
install.packages("randomForest")
install.packages("e1071")

# Carregando os pacotes
library(Amelia)
library(ggplot2)
library(caret)
library(reshape)
library(randomForest)
library(dplyr)
library(e1071)

# Carregando os datasets
dataset <- read.csv("credit-card.csv") → carrega os bibliotecas
# Visualizando os dados e sua estrutura
View(dataset) → visualizo todo o dataset
str(dataset) → visualizo os atributos
head(dataset) → podemos entender do dataset
##### Transformando e Limpando os Dados #####
# Convertendo os atributos idade, sexo, escolaridade e estado civil para fatores (categorias)

# Idade
head(dataset$AGE) → lista tabela/coluna idade
dataset$AGE <- cut(dataset$AGE, c(0,30,50,100), labels = c("Jovem","Adulto","Idoso")) → determina umas labels para
head(dataset$AGE)                               código idade

# Sexo
dataset$SEX <- cut(dataset$SEX, c(0,1,2), labels = c("Masculino","Feminino"))
head(dataset$SEX)

# Escolaridade
dataset$EDUCATION <- cut(dataset$EDUCATION, c(0,1,2,3,4),
                           labels = c("Pos Graduado","Graduado","Ensino Medio","Outros"))
head(dataset$EDUCATION)

# Estado Civil
dataset$MARRIAGE <- cut(dataset$MARRIAGE, c(-1,0,1,2,3),
                           labels = c("Desconhecido","Casado","Solteiro","Outros"))
head(dataset$MARRIAGE)

# Convertendo a variável que indica pagamentos para o tipo fator
dataset$PAY_0 <- as.factor(dataset$PAY_0)
dataset$PAY_2 <- as.factor(dataset$PAY_2)
dataset$PAY_3 <- as.factor(dataset$PAY_3)
dataset$PAY_4 <- as.factor(dataset$PAY_4)
dataset$PAY_5 <- as.factor(dataset$PAY_5)
dataset$PAY_6 <- as.factor(dataset$PAY_6)

# Alterando a variável dependente para o tipo fator
dataset$default.payment.next.month <- as.factor(dataset$default.payment.next.month)
head(dataset)
str(dataset)

# Renomeando a coluna de classe → lista todas colunas
colnames(dataset) → altera o nome da coluna
colnames(dataset)[25] <- "inadimplente"
colnames(dataset)

# Verificando valores missing e removendo do dataset
sapply(dataset, function(x) sum(is.na(x)))
missmap(dataset, main = "Valores Missing Observados")
dataset <- na.omit(dataset) → cria um mapa dos valores missing

# Removendo a primeira coluna ID
dataset$ID <- NULL → não traz nenhuma informação para a análise

# Total de inadimplentes versus não-inadimplentes
table(dataset$inadimplente) → mostra todos inadimplente e não-inadimplente

# Plot da distribuição usando ggplot
ggplot(inadimplente, data = dataset, geom = "bar") + theme(axis.text.x = element_text(angle = 90, hjust = 1))

# Set the seed
set.seed(12345)

# Amostragem estratificada. Selecione as linhas de acordo com a variável inadimplente como strata
TrainingDataIndex <- createDataPartition(dataset$inadimplente, p = 0.45, list = FALSE)
TrainingDataIndex

# Criar Dados de Treinamento como subconjunto do conjunto de dados com números de índice de linha
# conforme identificado acima e todas as colunas

```

```

trainData <- dataset[TrainingDataIndex,]
table(trainData$inadimplente)

# Veja porcentagens entre as classes
prop.table(table(trainData$inadimplente))

# Numero de linhas no dataset de treinamento
nrow(trainData)

# Compara as porcentagens entre as classes de treinamento e dados originais
DistributionCompare <- cbind(prop.table(table(trainData$inadimplente)), prop.table(table(dataset$inadimplente)))
colnames(DistributionCompare) <- c("Treinamento", "Original")
DistributionCompare

# Melt Data - Converte colunas em linhas
meltedDComp <- melt(DistributionCompare)
meltedDComp

# Plot para ver a distribuicao do treinamento vs original - eh representativo ou existe sobre / sob amostragem?
ggplot(meltedDComp, aes(x = X1, y = value)) + geom_bar(aes(fill = X2), stat = "identity", position = "dodge") + theme(axis.text.x =
element_text(angle = 90, hjust = 1))

# Tudo o que nao esta no dataset de treinamento esta no dataset de teste. Observe o sinal - (menos)
testData <- dataset[-TrainingDataIndex,]

# Usaremos uma validacao cruzada de 10 folds para treinar e avaliar modelo
TrainingParameters <- trainControl(method = "cv", number = 10)

Random Forest Classification Model

# Construindo o Modelo
rf_model <- randomForest(inadimplente ~ ., data = trainData)
rf_model

# Conferindo o erro do modelo
plot(rf_model, ylim = c(0,0.36))
legend('topright', colnames(rf_model$err.rate), col = 1:3, fill = 1:3)

# Importancia das variaveis preditoras para as previsoes
varImpPlot(rf_model)

# Obtendo as variaveis mais importantes
importance <- importance(rf_model)
varImportance <- data.frame(Variables = row.names(importance), Importance = round(importance[, 'MeanDecreaseGini'],2))

# Criando o rank de variaveis baseado na importancia
rankImportance <- varImportance %>%
  mutate(Rank = paste0('#', dense_rank(desc(Importance)))) %>%
  arrange(-Rank) %>%
  select(-Rank)

# Usando ggplot2 para visualizar a importancia relativa das variaveis
ggplot(rankImportance, aes(x = reorder(Variables, Importance), y = Importance, fill = Importance)) +
  geom_bar(stat='identity') +
  geom_text(aes(x = Variables, y = 0.5, label = Rank), hjust=0, vjust=0.55, size = 4, colour = 'red') +
  labs(x = 'Variables') +
  coord_flip()

# Previsoes
predictionrf <- predict(rf_model, testData)

# Confusion Matrix
cmrf <- confusionMatrix(predictionrf, testData$inadimplente, positive = "1")
cmrf

# Salvando o modelo
saveRDS(rf_model, file = "rf_model.rds")

# Carregando o modelo
modelo <- readRDS("rf_model.rds")

# Calculando Precision, Recall e F1-Score, que sao metricas de avaliacao do modelo preditivo
y <- testData$inadimplente
predictions <- predictionrf

precision <- posPredValue(predictions, y)
precision

recall <- sensitivity(predictions, y)
recall

F1 <- (2 * precision * recall) / (precision + recall)
F1

```

Data set → favorável → todos variáveis
 inadimplente → Ponto indica que estão usando todas as variáveis
 → todos variáveis → todos variáveis
 → todos variáveis → todos variáveis

importante

Estatística

Probabilística = estudo da aleatoriedade e incerteza

Estatística Descritiva = estrutura da

Estatística inferencial = População → amostra

Descrição = População

w.w.w. Daxpatterns.com/statistical-patterns/

Lo scripts de estatística

R

↳ Session

↳ Set working directory

↳ Choose directory

```

1 # Medidas de Posição
2 # Definindo a pasta de trabalho
3 # Substitua o caminho abaixo pela pasta no seu computador
4
5 setwd("D:/Dropbox/DSA/PowerBI-DataScience/Cap11/01-Medidas-Posicao") → Set Working D.
6 getwd() → mostra diretório → choose D.
7
8
9 # Carregando o dataset
10 vendas <- read.csv("Vendas.csv", fileEncoding = "windows-1252")
11
12 # Resumo do dataset
13 View(vendas) → mostra tabela
14 str(vendas) → Resumo dos dados
15 summary(vendas$Valor) → mínima | 1= 14 | media | mediana | 3= 17 | max
16 summary(vendas$Custo)
17
18 # Help = obtém a documentação da função
19
20 # Média
21 ?mean → média = average
22 mean(vendas$Valor) → ordena e conta o More than
23 mean(vendas$Custo) → ordema e conta o More than
24 mean(vendas$Valor, trim = 0.1) → Elmina valores missing → Ctrl + Enter
25 mean(vendas$Valor, na.rm = TRUE) → Elmina valores missing → excuta linhas
26
27
28 # Média Ponderada
29 ?weighted.mean
30 weighted.mean(vendas$Valor, w = vendas$Custo)
31 → médio ponderado → Peso
32
33 # Mediana
34 median(vendas$Valor)
35 median(vendas$Custo)
36
37 # Moda
38 # Criando uma função
39 getmode <- function(v) { → nome da função getmode
40 uniqv <- unique(v)
41 uniqv[which.max(tabulate(match(v, unqv)))] → parâmetro de entrada
42 }
43
44
45 # Obtendo a Moda
46 result <- getmode(vendas$Valor)
47 print(result)
48
49
50 # Criando gráfico de médias por Estado com ggplot2
51 install.packages("ggplot2")
52 library(ggplot2)
53 ggplot(vendas) + stat_summary(aes(x = vendas$Estado, y = vendas$Valor),
54 fun.y = mean, geom = "bar",
55 fill = "lightgreen", col = "grey50")

```

R = SAS = SPSS

Ferramentas de Estatística

Desvio padrão é o desvio médio da média

Coeficiente de Variação

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

s = desvio Padrão

x = médio

Posição Relativa

Verificam se um valor comparece com os outros

Percentil

Se compara a minha nota com a dos meus colegas da classe ou da turma
ideia da minha posição no Sólo

Percentil \neq Percentagem

Aluno tirou uma nota 36/45 teve um aproveitamento de 80% porém um percentil de 97%.
Significa que ele foi melhor que 96% da Sólo

Quantil

calcular (Average (vendas[Valor]), Filter(vendas, vendas[^{Unit}])
>= Percentil.exc (vendas[Valor], 0.8)))

Percentil.exc = Excluir o valor até 0.80%
ou seja retorna 20% da média do
Vendas [Valor]

Visualizations	
Values	Value
Valor	Média
Soma	
minimum	
maximum	
contagem	
desvioPadrao	
mediana	

Power Bi
Ja calcula Valores
já inclui os para
Padrão
Nós precisamos
média nova
Por contagem do
total Geral

Valor = count
► Show Valor as: Percent

```

1 > getwd()      Exibe caminho do arquivo
2 [1] "E:/Givaldo/DATA SCIENCE/CURSOS/Data Science
3 academy/PowerBI-DataScience-master/Cap11/05-Frequencia"
4 > dados <- read.table("usuarios.csv", dec = ".", sep = ",", h = T, fileEncoding =
5 "windows-1252")  começo os dados
6 > names(dados)           "N"          "estado_civil"  Exibe nome das
7 [1] "X"                  "N"          "salario"        colunas
8 [4] "grau_instrucao"    "n_filhos"   "idade_meses" 
9 [7] "idade_anos"        "idade_meses" "reg_procedencia"
10
11 > str(dados)
12 'data.frame': 36 obs. of 9 variables:
13   $ X : int 1 2 3 4 5 6 7 8 9 10 ...
14   $ N : int 1 2 3 4 5 6 7 8 9 10 ...
15   $ estado_civil : Factor w/ 2 levels "casado", "solteiro": 2 1 1 2 2 1 2 2 1 2 ...
16   $ grau_instrucao : Factor w/ 3 levels "ensino fundamental", ...: 1 1 1 2 1 1 1 2
17   2 ...
18   $ n_filhos : int NA 1 2 NA NA 0 NA NA 1 NA ...
19   $ salario  Float: num 4 4.56 5.25 5.73 6.26 6.66 6.86 7.39 7.59 7.44 ...
20   $ idade_anos : int 26 32 36 20 40 28 41 43 34 23 ...
21   $ idade_meses : int 3 10 5 10 7 0 0 4 10 6 ...
22   $ reg_procedencia: Factor w/ 3 levels "capital", "interior", ...: 2 1 1 3 3 2 2 1 1
23 > summary(dados$salario) Você pode fazer o sumário diretamente pelo
24 Min. 1st Qu. Median Mean 3rd Qu. Max.  nome da coluna
25 4.000 7.553 10.165 11.122 14.060 23.300
26
27 > freq <- table(dados$grau_instrucao)
28 > freq tabela de frequencia
29
30
31 ensino fundamental      ensino medio
32           12                  18
33 >
34 # Tabela de Frequências Relativas
35
36 > freq_rel <- prop.table(freq)
37
38 > freq_rel
39
40 ensino fundamental      ensino medio      superior
41           0.3333333       0.5000000     0.1666667
42 >
43 # Porcentagem (100 * freq_rel_table)
44
45 > p_freq_rel <- 100 * prop.table(freq_rel)
46
47 > p_freq_rel
48
49 ensino fundamental      ensino medio      superior
50           33.33333       50.00000     16.66667
51 > Possui o vetor de frequencia e soma a frequ.
52 # Adiciona linhas de total
53 > freq <- c(freq, sum(freq)) insira no variável = freq.
54 > freq_rel <- c(freq_rel, sum(freq_rel))
55 > p_freq_rel <- c(p_freq_rel, sum(p_freq_rel))
56 > names(freq)[4] <- "Total" → Cria uma 4ª coluna para gravar o total
57
58 # Tabela final
59 > tabela_final <- cbind(freq,
60 + freq_rel = round(freq_rel, digits = 2),
61 + p_freq_rel = round(p_freq_rel, digits = 2))
62 > tabela_final
63
64 ensino fundamental      freq freq_rel p_freq_rel
65 ensino medio             12     0.33     33.33
66                           18     0.50     50.00

```

Exibe caminho do arquivo

Exibe nome das colunas

começo os dados

Você pode fazer o sumário diretamente pelo nome da coluna

12 categorias ensino fundamental

18 categorias ensino médio

superior

→ Possui o vetor de frequencia e soma a frequ.

→ Adiciona linhas de total

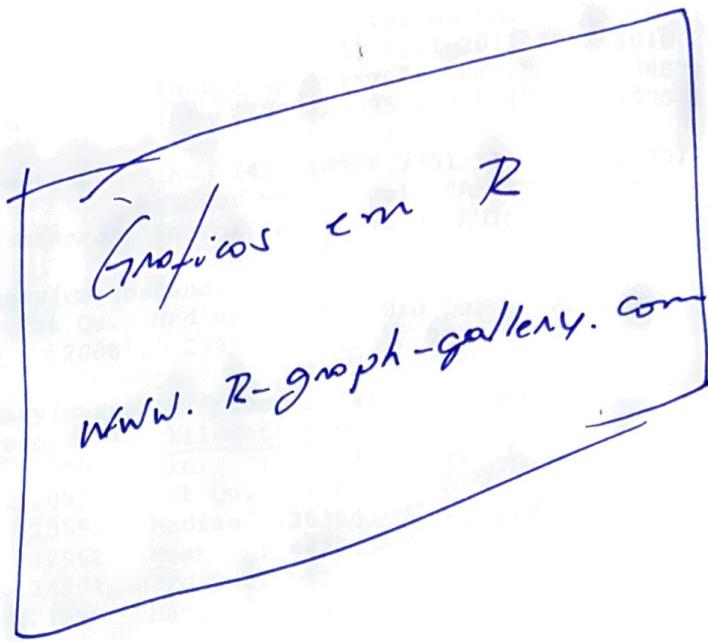
→ insira no variável = freq.

→ Cria uma 4ª coluna para gravar o total

→ arredonda o valor

Faz a uniao das colunas

66 superior 6 0.17 16.67
67 Total 36 1.00 100.00
68 >



```

1
2 head(carros)
3     ano modelo preco kilometragem      cor transmissao
4 1 2011    SEL 21992          7413 Bege      AUTO
5 2 2011    SEL 20995          10926 Cinza      AUTO
6 3 2011    SEL 19995          7351 Prata      AUTO
7 4 2011    SEL 17809          11613 Cinza      AUTO
8 5 2012    SE 17500           8367 Branco     AUTO
9 6 2010    SEL 17495          25125 Prata      AUTO
10
11 > str(carros)
12 'data.frame': 150 obs. of 6 variables:
13   $ ano      : int 2011 2011 2011 2011 2012 2010 2011 2010 2011 2010 ...
14   $ modelo   : Factor w/ 3 levels "SE","SEL","SES": 2 2 2 2 1 2 2 2 3 3 ...
15   $ preco    : int 21992 20995 19995 17809 17500 17495 17000 16995 16995 16995
16   ...
17   $ kilometragem: int 7413 10926 7351 11613 8367 25125 27393 21026 32655 36116 ...
18   $ cor       : Factor w/ 9 levels "Azul","Bege",...: 2 4 6 4 3 6 1 6 6 6 ...
19   $ transmissao: Factor w/ 2 levels "AUTO","MANUAL": 1 1 1 1 1 1 1 1 1 1 ...
20
21 > summary(carros$ano)
22   Min. 1st Qu. Median Mean 3rd Qu. Max.
23   2000    2008    2009  2009    2010    2012
24
25 > summary(carros[c('preco', 'kilometragem')])
26   preco      kilometragem
27   Min. : 3800  Min. : 4867
28   1st Qu.:10995 1st Qu.: 27200
29   Median :13592 Median : 36385
30   Mean   :12962 Mean   : 44261
31   3rd Qu.:14904 3rd Qu.: 55125
32   Max.   :21992  Max.   :151479
33
34 > mean(carros$preco)               media
35 [1] 12961.93
36
37 > median(carros$preco)            mediana
38 [1] 13591.5
39
40 > quantile(carros$preco, probs = c(0.01, 0.99))
41   1%    99%
42   5428.69 20505.00
43
44 > quantile(carros$preco, seq(from = 0, to = 1, by = 0.20))           De 20% em 20%
45   0%    20%    40%    60%    80%    100%
46   3800.0 10759.4 12993.8 13992.0 14999.0 21992.0
47
48 > IQR(carros$preco) # Diferença entre Q3 e Q1
49 [1] 3909.5
50
51 > range(carros$preco)            Min e Max
52 [1] 3800 21992
53
54 > summary(carros$preco)
55   Min. 1st Qu. Median Mean 3rd Qu. Max.
56   3800 10995 13592 12962 14904 21992
57
58 > diff(range(carros$preco))
59 [1] 18192

```

Retorno Sumário das duas colunas indicadas

```

1 # Medidas de Posição Relativa
2
3 # Definindo a pasta de trabalho
4 # Substitua o caminho abaixo pela pasta no seu computador
5 setwd("D:/Dropbox/DSA/PowerBI-DataScience/Cap11/04-Medidas-Posicao-Relativa")
6 getwd()
7
8
9 # Carregando o dataset
10 carros <- read.csv("carros.csv")
11
12 # Resumo dos dados
13 head(carros)
14 str(carros)
15
16 # Medidas de Tendência Central
17 summary(carros$ano)
18 summary(carros[c('preco', 'kilometragem')]) # a função "c" é um vetor, aqui
mostra o sumario de colunas específicas
19
20 ## Explorando variáveis numéricas
21 mean(carros$preco)
22
23 median(carros$preco)
24
25
26 quantile(carros$preco)
27
28
29 quantile(carros$preco, probs = c(0.01, 0.99))
30
31
32 quantile(carros$preco, seq(from = 0, to = 1, by = 0.20))
33
34
35 IQR(carros$preco) # Diferença entre Q3 e Q1
36
37
38 range(carros$preco)
39
40
41 summary(carros$preco)
42
43
44 diff(range(carros$preco))
45

```

```

1 setwd("E:/Givaldo/DATA SCIENCE/CURSOS/Data Science
2 academy/PowerBI-DataScience-master/Cap11/06-Plots")
3 > my_vector = c(3,12,5,18,45)
4 > names(my_vector) = c("A","B","C","D","E")
5 > my_vector
6 A B C D E
7 3 12 5 18 45
8 > barplot(my_vector)
9 > barplot(my_vector, col = c(1,2,3,4,5) )
10 > png("barplot.png", width = 400, height = 400 )
11 > barplot(my_vector, col = rgb(0.5,0.1,0.6,0.0), xlab = "Categorias", ylab =
12 "Valores", main = "Barplot em R", ylim = c(0,60) )
13 > dev.off()
14 RStudioGD
15 2
16 > library(ggplot2)
17 Warning message:
18 package 'ggplot2' was built under R version 3.5.2
19 > head(mtcars)
20          mpg cyl disp  hp drat    wt  qsec vs am gear carb
21 Mazda RX4     21.0   6 160 110 3.90 2.620 16.46  0  1    4    4
22 Mazda RX4 Wag 21.0   6 160 110 3.90 2.875 17.02  0  1    4    4
23 Datsun 710    22.8   4 108 93 3.85 2.320 18.61  1  1    4    1
24 Hornet 4 Drive 21.4   6 258 110 3.08 3.215 19.44  1  0    3    1
25 Hornet Sportabout 18.7   8 360 175 3.15 3.440 17.02  0  0    3    2
26 Valiant       18.1   6 225 105 2.76 3.460 20.22  1  0    3    1
27 > ggplot(mtcars, aes(x=as.factor(cyl)) +
28 + geom_bar())
29 > ggplot(mtcars, aes(x=as.factor(cyl), fill=as.factor(cyl)) +
30 + geom_bar() +
31 + scale_fill_manual(values = c("red", "green", "blue") )
32 > data = data.frame(group = c("A","B","C","D"), value=c(33,62,56,67) )
33 > data = data.frame(group = c("A","B","C","D"), value=c(33,62,56,67) )
34 > # Barplot
35 > ggplot(data, aes(x = group, y = value, fill = group)) +
36 + geom_bar(width = 0.85, stat="identity") Grafico de barras
37 > # Pie Chart
38 > slices <- c(10, 12, 4, 16, 8)
39 > lbls <- c("US", "UK", "Australia", "Germany", "France") Cria um pie chart
40 > pie(slices, labels = lbls, main = "Beer per Country")
41 >
42 >
43 > # Pie Chart com percentuais
44 > slices <- c(10, 12, 4, 16, 8)
45 > lbls <- c("US", "UK", "Australia", "Germany", "France")
46 > pct <- round(slices/sum(slices)*100)
47 > lbls <- paste(lbls, pct)
48 > lbls <- paste(lbls,"%",sep="")
49 > pie(slices, labels = lbls, col=rainbow(length(lbls)),
50 + main="Beer per Country") instala pacote
51 >
52 >
53 > # Pie Chart 3D
54 > install.packages("plotrix") instala pacote
55 Installing package into 'C:/Users/gival/Downloads/R/win-library/3.5'
56 (as 'lib' is unspecified)
57 trying URL 'https://cran.rstudio.com/bin/windows/contrib/3.5/plotrix_3.7-4.zip'
58 Content type 'application/zip' length 1054471 bytes (1.0 MB)
59 downloaded 1.0 MB
60
61 package 'plotrix' successfully unpacked and MD5 sums checked
62
63 The downloaded binary packages are in
64
65 C:\Users\Public\Documents\Wondershare\CreatorTemp\RtmpWMzVqJ\downloaded_package

```

```

66 Warning message:
67 package 'plotrix' was built under R version 3.5.1
68
69 > slices <- c(10, 17, 4, 14, 9)
70 > lbls <- c("US", "UK", "Australia", "Germany", "France")
71 > pie3D(slices, labels=lbls, explode=0.1,
72 +         main="Beer per Country")
73 >
74 >
75 > # Line
76 >
77 > # Dados
78 > cars <- c(1, 3, 6, 4, 9) Cria vetores consiste trucks
79 > trucks <- c(2, 5, 4, 5, 12)
80 >
81 > # Plot
82 > plot(cars, type="o", col="blue", ylim=c(0,12)) Plata Grafico de
83 > lines(trucks, type="o", pch=22, lty=2, col="red") linhos
84 >
85 > title(main="Autos", col.main="red", font.main=4) Coloreia ticks "Autos"
86 >
87 >
88 > # Boxplot
89 > library(ggplot2) Chama a biblioteca ggplot2
90 > head(mpg)
91 >
92 # A tibble: 6 x 11
93   manufacturer model displ year cyl trans drv cty hwy fl
94   <chr> <chr> <dbl> <int> <chr> <chr> <int> <int> <chr>
95  1 audi     a4      1.8  1999     4 auto~ f     18    29 p
96  2 audi     a4      1.8  1999     4 manu~ f     21    29 p
97  3 audi     a4      2     2008     4 manu~ f     20    31 p
98  4 audi     a4      2     2008     4 auto~ f     21    30 p
99  5 audi     a4      2.8  1999     6 auto~ f     16    26 p
100 6 audi     a4      2.8  1999     6 manu~ f     18    26 p
101 # ... with 1 more variable: class <chr>
102 >
103 > # Plot
104 > ggplot(mpg, aes(x=reorder(class, hwy), y=hwy, fill=class)) +
105 +   geom_boxplot() +
106 +   xlab("class") +
107 +   theme(legend.position="none")
108 >
109 >
110 > # Scatter Plot
111 > library(ggplot2)
112 > data = data.frame(cond = rep(c("condition_1", "condition_2"), each=10),
113 +                     my_x = 1:100 + rnorm(100, sd=9), my_y = 1:100 +
114 +                     rnorm(100, sd=16) )
115 >
116 > ggplot(data, aes(x=my_x, y=my_y)) +
117 +   geom_point(shape=1) → Cria um grafico de
118 > # Adiciona linha de regressao dispersao
119 > ggplot(data, aes(x=my_x, y=my_y)) +
120 +   geom_point(shape=1) +
121 +   geom_smooth(method = lm, color="red", se=FALSE)
122 >
123 > # Adiciona smooth cua linha
124 > ggplot(data, aes(x=my_x, y=my_y)) +
125 +   geom_point(shape=1) +
126 +   geom_smooth(method=lm, color="red", se=TRUE)
127 >
128 >
129 > # Treemap
130 > install.packages("treemap") instala o pacote treemap
131 Installing package into 'C:/Users/gival/Documents/R/win-library/3.5'
132 (as 'lib' is unspecified)
133 also installing the dependencies 'httpuv', 'mime', 'jsonlite', 'xtable',

```

```

134
135     There is a binary version available but the source version is
136     later:
137         binary source needs_compilation
138             TRUE
139     later 0.7.5 0.8.0
140
141 package 'httpuv' successfully unpacked and MD5 sums checked
142 package 'mime' successfully unpacked and MD5 sums checked
143 package 'jsonlite' successfully unpacked and MD5 sums checked
144 package 'xtable' successfully unpacked and MD5 sums checked
145 package 'ht倾ttools' successfully unpacked and MD5 sums checked
146 package 'sourcetools' successfully unpacked and MD5 sums checked
147 package 'promises' successfully unpacked and MD5 sums checked
148 package 'gridBase' successfully unpacked and MD5 sums checked
149 package 'igraph' successfully unpacked and MD5 sums checked
150 package 'shiny' successfully unpacked and MD5 sums checked
151 package 'treemap' successfully unpacked and MD5 sums checked
152
153 The downloaded binary packages are in
154
155
156     C:\Users\Public\Documents\Wondershare\CreatorTemp\RtmpWMzVqJ\downloaded_package
157         s
158     installing the source package 'later'
159
160     trying URL 'https://cran.rstudio.com/src/contrib/later_0.8.0.tar.gz'
161     Content type 'application/x-gzip' length 40237 bytes (39 KB)
162     downloaded 39 KB
163
164     * installing *source* package 'later' ...
165     ** package 'later' successfully unpacked and MD5 sums checked
166     ** libs
167
168     ** R
169     ** inst
170     ** byte-compile and prepare package for lazy loading
171     ** help
172     *** installing help indices
173     converting help for package 'later'
174     finding HTML links ... done
175         later                      html
176         loop_empty                 html
177         next_op_secs               html
178         run_now                   html
179
180     ** building package indices
181     ** installing vignettes
182     ** testing if installed package can be loaded
183     *** arch - i386
184     *** arch - x64
185     * DONE (later)
186     In R CMD INSTALL
187
188 The downloaded source packages are in
189
190
191     'C:\Users\Public\Documents\Wondershare\CreatorTemp\RtmpWMzVqJ\downloaded_packages'
192
193 > library(treemap)
194 Warning message:
195 package 'treemap' was built under R version 3.5.2
196
197 > # Dados
198 > group=c(rep("group-1",4),rep("group-2",2),rep("group-3",3))
199 > subgroup= paste("subgroup", c(1,2,3,4,1,2,1,2,3), sep="-")
200 > value=c(13,5,22,12,11,7,3,1,23)
201 > data=data.frame(group,subgroup,value)
202 >
203 > # Labels

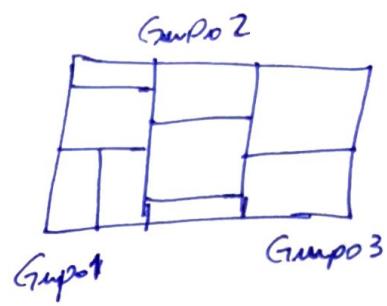
```

Grandes dados
factios

```

198 > treemap(data, index=c("group", "subgroup"),
199   + vSize="value", type="index",
200   + fontsize.labels=c(15,10),
201   + fontcolor.labels=c("white", "orange"),
202   + fontface.labels=c(0,1),
203   + bg.labels=c("transparent"),
204   + align.labels=list(
205     +   c("center", "center"),
206     +   c("right", "bottom")),
207   + overlap.labels=0.5,
208   + inflate.labels=F,
209   +
210 >
211 > # Customizando
212 > treemap(data, index=c("group", "subgroup"), vSize="value", type="index",
213   + border.col=c("black", "white"),
214   + border.lwds=c(1,2)
215 >
216 >
217 >
218 > # Histograma
219 > x <- mtcars$mpg
220 > 
221 >
222 > h <- hist(x, breaks = 10, col="red", xlab = "Miles Per Gallon",
223   + main = "Histograma com Curva de Distribuicao")
224 >
225 > xfit <- seq(min(x),max(x),length=40)
226 > yfit <- dnorm(xfit,mean=mean(x),sd=sd(x))
227 > yfit <- yfit*diff(h$mids[1:2])*length(x) 
228 > lines(xfit, yfit, col="blue", lwd=2)
229 >
230 > # Usando o ggplot2
231 > library(ggplot2)
232 >
233 > # dataset
234 > data = data.frame(value = rnorm(10000))
235 >
236 > # Custom Binning. I can just give the size of the bin
237 > ggplot(data, aes(x=value)) +
238   + geom_histogram(binwidth = 0.05)
239 >
240 > # Uniform color
241 > ggplot(data, aes(x=value)) +
242   + geom_histogram(binwidth = 0.2, color="white", fill=rgb(0.2,0.7,0.1,0.4) )
243 >
244 > # Proportional color
245 > ggplot(data, aes(x=value)) +
246   + geom_histogram(binwidth = 0.2, aes(fill = ..count..) )
247 >

```



Group 1

Group 2

Group 3

Group 4

Cria a curva de
distribuicao.