

Introdução a Ciência de Dados



Professor: Alex Pereira

SQL é o sabre de luz de um Cientista de Dados



Query de pessoas que não tomaram a 2ª dose da vacina do COVID-19 (1)

- Selecionar **pessoas** que ainda não tomaram a 2ª dose
 - Esta é uma tabela de transação
 - ✓ e não de entidades

```
SELECT distinct(v.id_paciente)
FROM `basedosdados.br_ms_vacinacao_covid19.microdados_vacinacao` v
WHERE v.vacina != '88'
GROUP BY v.id_paciente
HAVING min(v.data_aplicacao) = max(v.data_aplicacao)
LIMIT 10
```

* O LIMIT é opcional (recomendável) para testes

Query de pessoas que não tomaram a 2ª dose da vacina do COVID-19 (2) – Sub-select

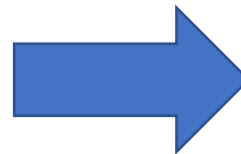
```
select sum( CASE
    WHEN vagg.vacina='86' THEN
        case when (DATE_DIFF(current_date, vagg.data_aplicacao, day) > 30) then 1 else 0 end
    ELSE
        case when (DATE_DIFF(current_date, vagg.data_aplicacao, day) > 90) then 1 else 0 end
    END ) AS naotomou2aDose
    , vagg.uf, vagg.vacina
from `basedosdados.br_ms_vacinacao_covid19.microdados_vacinacao` vagg
where vagg.id_paciente in
( SELECT distinct(v.id_paciente)
  FROM `basedosdados.br_ms_vacinacao_covid19.microdados_vacinacao` v
  WHERE v.vacina != '88'
  GROUP BY v.id_paciente
  HAVING min(v.data_aplicacao) = max(v.data_aplicacao) )
group by uf, vacina;
```

Pivot Table (Tabela Dinâmica)

- Operação de manipulação de dados realizada em vários softwares
 - SQL, MS Excel, Google Sheets, Data Studio, Pandas, entre outros
- Transforma dados dispostos em linhas
 - Para o equivalente na forma de colunas

	Country	Quarter	Year	Revenue ...
1.	United States	Q3	2017	\$198.51
2.	Hong Kong	Q3	2017	\$70.21
3.	Canada	Q3	2017	\$52.74
4.	Mexico	Q3	2017	\$42.36
5.	United States	Q4	2017	\$36.37
6.	Venezuela	Q3	2017	\$33.49
7.	Ireland	Q4	2017	\$27.25
8.	Australia	Q3	2017	\$27.10

Pivot



	2017	
Country	Q3	Q4
United States	\$198.51	\$36.37
Hong Kong	\$70.21	-
Canada	\$52.74	\$6.42
Venezuela	\$33.49	-
Australia	\$27.10	\$15.70
Mexico	\$42.36	-

Pivot Table no Google Sheets

- Copie para uma tabela sua, selecione o intervalo de células e
 - Clique em Data -> Pivot table (ou Dados - > Tabela dinâmica)

Country	Quarter	Year	Revenue	SUM of Revenue by Quarter		
Country	Q3	Q4	Grand Total			
United States	Q3	2017	198.5	(not set)	23.8	23.8
Hong Kong	Q3	2017	70.21	Australia	27.1	27.1
Canada	Q3	2017	52.74	Canada	52.74	52.74
Mexico	Q3	2017	42.36	Hong Kong	70.21	70.21
United States	Q4	2017	36.37	Ireland	27.25	27.25
Venezuela	Q3	2017	33.49	Mexico	42.36	42.36
Ireland	Q4	2017	27.25	Singapore	25.19	25.19
Australia	Q3	2017	27.1	United States	198.5	234.87
Singapore	Q3	2017	25.19	Venezuela	33.49	33.49
(not set)	Q4	2017	23.8	Grand Total	449.59	537.01

Rows Add

Country ×

Order Ascending Sort by Country

☒ Show totals

Columns Add

Quarter ×

Order Ascending Sort by Quarter

☒ Show totals

Values Add

Revenue ×

Summarize by SUM Show as Default

Google Sheet:

<https://docs.google.com/spreadsheets/d/14hQYqRViuvd7EC1I-SHCCyvO97k8NkNppMcl62HeSM0/edit?usp=sharing>

XLSX:

<https://docs.google.com/spreadsheets/d/16B0uAFH1hubqvXytaxEkh9J8fXAJugYi/edit?usp=sharing>

Reshaping / Pivoting (Pivotar)

- Método pivot

- 3 argumentos: **index**, **columns**, **values**

- ✓ `df.pivot(index='Aluno', columns='Disciplina', values='Objetiva')`

- a função `melt()` faz a operação de despivotar

	Aluno	Disciplina	Objetiva	Discursiva
0	AlunoA	Portugues	8.5	6
1	AlunoA	Matematica	7.5	6.5
2	AlunoB	Geografia	9	7.5
3	AlunoB	História	10	7

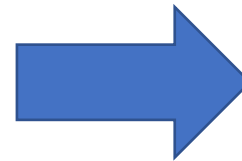
	Disciplina	Geografia	História	Matematica	Portugues
Aluno	AlunoA	NaN	NaN	7.5	8.5
	AlunoB	9	10	NaN	NaN

Pivotar

E quando houver valores repetidos ?

- Pivotar com o mesmo método pivot() gera exceção
 - Neste caso, use o método pivot_table
 - ✓ mean é a métrica padrão de cálculo sobre a de agregação

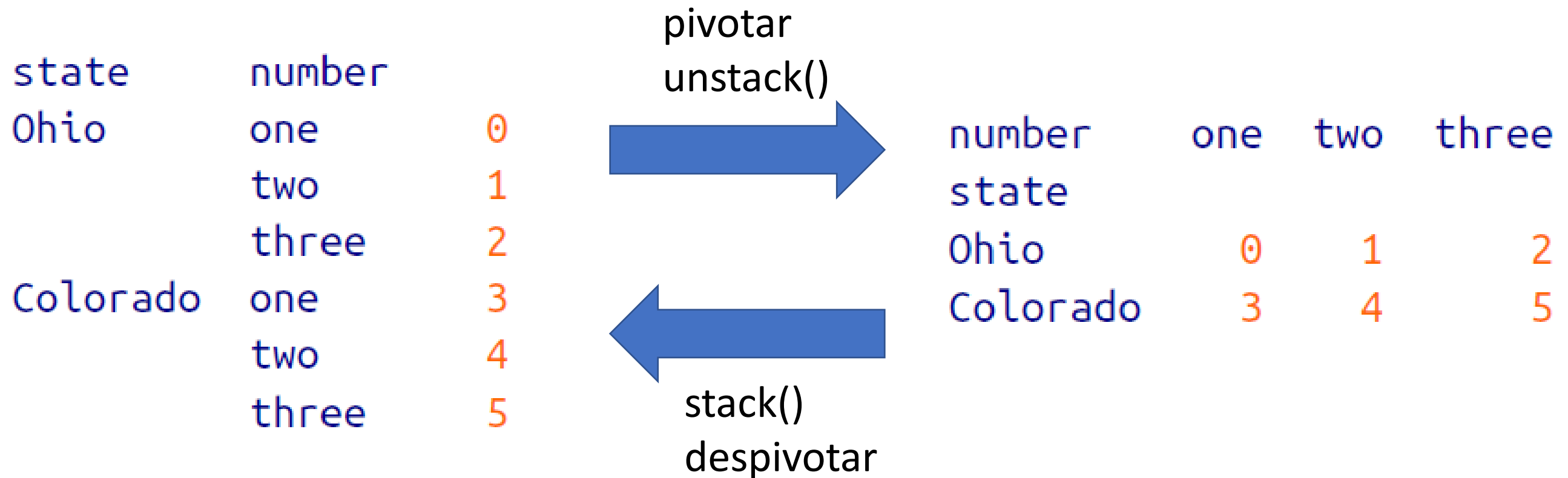
	Aluno	Disciplina	Objetiva	Discursiva
0	AlunoA	Portugues	8.5	6.0
1	AlunoA	Matematica	7.5	6.5
2	AlunoA	Geografia	9.0	7.5
3	AlunoA	Geografia	10.0	7.0
4	AlunoA	História	9.0	8.0
5	AlunoB	Portugues	8.5	8.5
6	AlunoB	Matematica	7.5	7.5
7	AlunoB	Geografia	9.0	9.0
8	AlunoB	História	10.0	10.0



Disciplina	Geografia	História	Matematica	Portugues
Aluno				
AlunoA	9.5	9.0	7.5	8.5
AlunoB	9.0	10.0	7.5	8.5

Reshaping / Pivoting com Índice Hierárquico

- Método stack/unstack (Pivotar com índice hierárquico)
 - stack = empilhar



Pivot Table no Pandas

- 2 Exemplos
 - Link para o caderno Colab de Teoria [aqui](#)
- Tempo para a Atividade 2.1: 10min
 - O dataset está na aba Orders (desta [planilha](#))
 - Link para o caderno Colab de Atividade [aqui](#)



Google Data Studio



<https://support.google.com/datastudio/?hl=pt-BR>

Dimensões e Métricas

- Dimensões são atributos usados para
 - descrever, segmentar/agrupar, organizar e ordenar dados
 - ✓ Data, idade, sexo, cidade, dispositivo, etc.
- Métricas são medidas quantitativas extraídas dos dados

Qual a equivalência com uma query SQL?

Device Category ?		metrics →	Users ?	New Users ?	Sessions ?	Bounce Rate ?	Pages / Session ?	Avg. Session Duration ?
			14,695 % of Total: 100.00% (14,695)	12,640 % of Total: 100.09% (12,629)	17,749 % of Total: 100.00% (17,749)	45.37% Avg for View: 45.37% (0.00%)	4.25 Avg for View: 4.25 (0.00%)	00:02:54 Avg for View: 00:02:54 (0.00%)
<input type="checkbox"/>	1. desktop		10,117 (69.49%)	8,520 (67.41%)	12,409 (69.91%)	42.99%	4.42	00:03:11
<input type="checkbox"/>	2. mobile		4,153 (28.53%)	3,851 (30.47%)	4,990 (28.11%)	50.48%	3.84	00:02:13
<input type="checkbox"/>	3. tablet		289 (1.99%)	269 (2.13%)	350 (1.97%)	56.86%	4.08	00:02:42

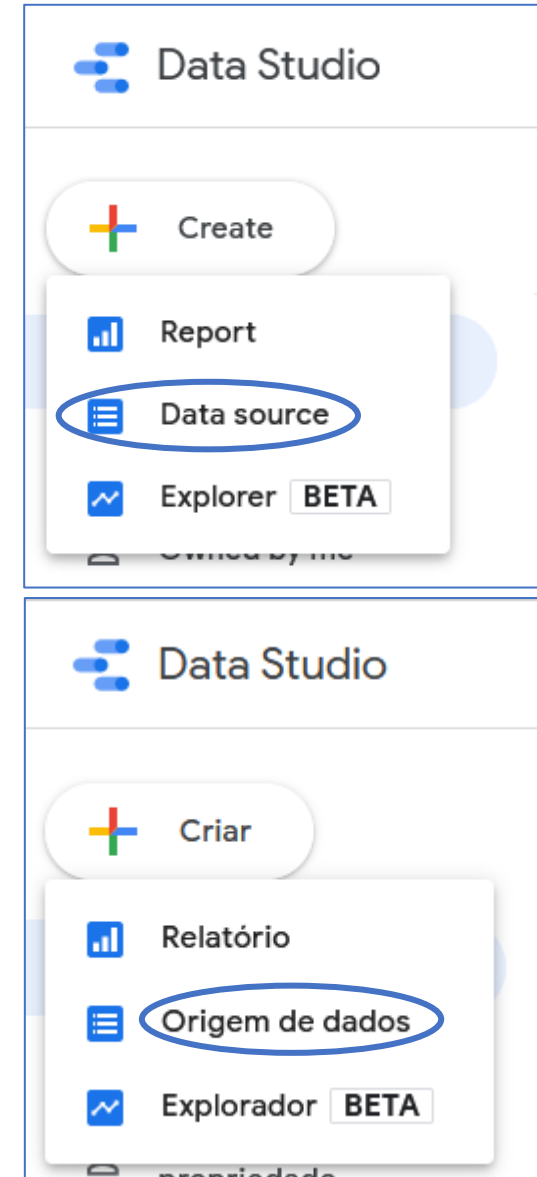
Data Connectors e Data Sources

- Conectores

- conectam o Data Studio aos seus dados
 - ✓ Existem conectores nativos/Google/gratuitos e de terceiros
 - Você também pode criar seu próprio conector.

- Data Sources (Origem de dados)

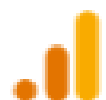
- Ao conectar seus dados, um objeto Data Source
 - ✓ será criado no seu Data Studio
 - Representam uma instância particular de um conector e uma base de dados
 - Uma maneira segura de filtrar dados ao compartilhá-los



Conectores do Google

Google Connectors (19)

Connectors built and supported by Data Studio [Learn more](#)



Google Analytics



By Google

Connect to Google Analytics.



Google Ads



By Google

Connect to Google Ads performance report data.



Google Sheets



By Google

Connect to Google Sheets.



BigQuery



By Google

Connect to BigQuery tables and custom queries.



File Upload



By Google

Connect to CSV (comma-separated values) files.



Campaign Manager 360



By Google

Connect to Campaign Manager 360 data.



Cloud Spanner



By Google

Connect to Google Cloud Spanner databases.



Cloud SQL for MySQL



By Google

Connect to Google Cloud SQL for MySQL databases.



Display & Video 360



By Google


Connect to Display & Video 360 report data.

Conector do Google Sheets

[← SELECT CONNECTOR](#)

4

[CONNECT](#)




Google Sheets


By Google



The Google Sheets connector allows you to access data stored in a Google Sheets worksheet.


[LEARN MORE](#)[REPORT AN ISSUE](#)

ALL ITEMS	Spreadsheet	Worksheet	Options
OWNED BY ME 1	pivot_table 2	revenue	<input checked="" type="checkbox"/> Use first row as headers
SHARED WITH ME	Backlog_Busca	Orders 3	<input checked="" type="checkbox"/> Include hidden and filtered cells
STARRED	dataset_tribo_sgd		Column headers must be unique. Columns with empty headers will not work.
URL	FormularioInscricaoPython_TriboCD_...		Optional Range, e.g. A1:B52
OPEN FROM GOOGLE DRIVE 	Nomes de Serviços #Gov360		
	[MBA IDP] Av. 1º bloco - notas		
	history (Responses)		
	datasus_morbidade		


Conector do Google Sheets

As visualizações da comunidade: [Ativado](#) | Edição dos campos nos relatórios: [Ativado](#)  **CRIAR RELATÓRIO** **EXPLORAR**





[←](#) **EDITAR CONEXÃO** | **FILTRAR POR E-MAIL**  **ADICIONAR UM CAMPO**  **ADICIONAR UM PARÂMETRO**

Campo ↓ **Tipo** ↓ **Agregação padrão** ↓  **Pesquis**

DIMENSÕES (14)

CategoryID	:	123	Número	▼	Soma	▼
CategoryName	:	ABC	Texto	▼	Nenhu	
CustomerID	:	123	Número	▼	Soma	
EmployeeID	:	123	Número	▼	Soma	
FirstName	:	ABC	Texto	▼	Nenhu	
LastName	:	ABC	Texto	▼	Nenhu	
OrderDate	:		Data	▼	Nenhu	

Tipo do Dado

- 123 Numérico ▶
- ABC Texto ▶
-  Data e hora ▶
-  Booleano ▶
-  Informações geográficas ▶
- 123 Moeda ▶
-  URL ▶

Campo Calculado (Calculated Field)

- Transformar, classificar ou fazer cálculos com seus dados
 - Cálculos matemáticos, manipulação de textos, datas, dados geográficos, lógica (Se Então, AND, OR)
- Criar novas métricas e dimensões derivadas dos seus dados
- O campo pode ser calculado para cada registro (linha)
 - de dados de um gráfico/tabela que o inclua
- Exemplo:

Fórmula (?) FORMATAR FÓRMULA

1 CONCAT (UPPER (LastName), ', ', UPPER (FirstName))

✓ CANCELAR SALVAR

Conector do Google Sheets

- Faça uma cópia desta [planilha](#), crie um data source
 - [Demonstração](#)
 - ✓ 5 min para criar o data source

← SELECT CONNECTOR

CONNECT



Google Sheets

By Google

The Google Sheets connector allows you to access data stored in a Google Sheets worksheet.

[LEARN MORE](#)

[REPORT AN ISSUE](#)

ALL ITEMS	Spreadsheet	Worksheet	Options
OWNED BY ME	pivot_table	revenue	<input checked="" type="checkbox"/> Use first row as headers
SHARED WITH ME	Backlog_Busca	Orders	<input checked="" type="checkbox"/> Include hidden and filtered cells
STARRED	dataset_tribo_sgd		Column headers must be unique. Columns with empty headers will not work
URL	FormularioInscricaoPython_TriboCD_...		Optional Range, e.g. A1:B52
OPEN FROM GOOGLE DRIVE	Nomes de Serviços #Gov360		
	[MBA IDP] Av. 1º bloco - notas		
	history (Responses)		
	datasus_morbidade		

Atividade 2.2 (5 min)

2.2) Calcular o Subtotal de vendas (em \$) por produto, e mostra-lo numa tabela

	ProductName	Record Count ▾	subtotal
1.	Raclette Courdavault	54	82.280
2.	Guaraná Fantástica	51	5.062,5
3.	Camembert Pierrot	51	53.618
4.	Gorgonzola Telino	51	17.462,5
5.	Gnocchi di nonna Alice	50	47.994
6.	Tarte au sucre	48	53.391,9
7.	Jack's New England Clam...	47	9.466,65
8.	Rhönbräu Klosterbier	46	8.951,25
9.	Chang	44	20.083
10.	Pavlova	43	20.207,1
1 - 77 / 77 < >			

Conceitos envolvidos

- Dimensões;
- Métricas; e
- Campos Calculados;

- Resolução e criação de Labels para o enunciado do exercício
 - Criar campo calculado, definir a dimensão e a métrica
- Quais operações foram realizadas para mostrar a coluna subtotal?

Atividade 2.3 (5 min)

2.3) Mostrar numa tabela o Subtotal de vendas (em \$) por categoria

⋮

	CategoryName	subtotal ▾
1.	Beverages	309.582,25
2.	Dairy Products	269.128,3
3.	Meat/Poultry	190.682,69
4.	Confections	190.328,54
5.	Seafood	149.059,53
6.	Condiments	122.343
7.	Produce	111.395
8.	Grains/Cereals	106.848

1 - 8 / 8 < >

- Resolução

- Definir a dimensão (Categoria)
- Definir o campo da métrica (subtotal)
 - ✓ Assegurar-se de que a função de agregação é a que você deseja

Atividade 2.4 (5 min)

2.4) Calcular a média de vendas (em \$) por categoria

⋮

	CategoryName	subtotal ▾
1.	Meat/Poultry	1.102,21
2.	Produce	819,08
3.	Beverages	766,29
4.	Dairy Products	735,32
5.	Confections	569,85
6.	Condiments	566,4
7.	Grains/Cereals	545,14
8.	Seafood	451,7

1 - 8 / 8 < >

- Resolução

- Definir a dimensão (Categoria)
- Definir o campo da métrica (subtotal)
 - ✓ Assegurar-se de que a função de agregação é a que você deseja

Atividade 2.5 (5 min)

2.5) Calcular a média de vendas (em \$) por categoria e o share de cada shipper em \$ (a parte que coube a cada transportador entregar)

⋮

	CategoryName	Shipper...	subtotal ▾
1.	Meat/Poultry	2	1.354,81
2.	Meat/Poultry	3	1.002,51
3.	Produce	3	934,31
4.	Beverages	3	903,21
5.	Beverages	2	834,39
6.	Meat/Poultry	1	818,02
7.	Produce	2	812,03
8.	Dairy Products	1	782,22

1 - 24 / 24 < >

- Resolução

- Definir as dimensões (Categoria e ShipperID)
- Definir o campo da métrica (subtotal)
 - ✓ Assegurar-se de que a função de agregação é a que você deseja

Funções do Google Data Studio

- REGEXP_MATCH
 - REGEXP_MATCH(CategoryName, 'Bev.* | Dai.*')
- COUNT
 - Conta registros
- DISTINCT_COUNT
 - Conta registros distintos
- SUM, AVG, MEDIAN, MIN, MAX, ABS
 - STDDEV, VARIANCE, PERCENTILE

REGEXP_MATCH(X, Reg Expr) ✕

Resumo

Retornará "true" se X corresponder a Y. Caso contrário, retornará "false".

X

Um campo ou uma expressão.

Funções do Google Data Studio

- IF

- IF(condition, true_result, false_result)

- ✓ IF(Price > 20, "Caro", "Barato")

- ✓ IF(REGEXP_MATCH(UF,'DF|GO|MT|MS')=TRUE, 'Centro-Oeste', 'Outro')

- CAST

- Serve para conversão de tipos

- ✓ CAST(Price AS TEXT)

CAST(X AS TYPE) ×

Resumo

Definir o campo ou a expressão para TIPO, onde TIPO pode ser NUMBER ou TEXT. Não é permitido usar campos agregados em CAST.

X AS TYPE

X é um campo ou uma expressão. TYPE pode ser NUMBER ou TEXT.

Função CASE

- Útil para classificar dados

- Exemplo 1

- ✓ CASE

- WHEN STARTS_WITH (CEP, '708') THEN 'DF'**

- ELSE 'Outro' END**

- Exemplo 2

- ✓ CASE

- WHEN REGEXP_MATCH (FirstName, 'Anne.*|Janet.*') THEN 'Equipe1'**

- WHEN REGEXP_MATCH (FirstName, 'Michael.*|Steven.*') THEN 'Equipe2'**

- ELSE 'Equipe3' END**

Filtro

- Serve para filtrar (não mostrar) linhas/registros
 - De um gráfico ou tabela
- Qual o equivalente no SQL ?
- Exemplos:
 - Filtrar registros com valores nulos
 - ✓ Regex
 - Filtrar categorias indesejadas
 - Filtrar outliers

Criar filtro

Nome: Nulos

Origem de dados: pivot_table - Orders

Excluir ABC FirstName

E

Esse filtro tem 1 cláusula

- Igual a (=)
- Contém
- Começa com
- Correspondência RegExp
- RegExp contém
- Em
- Nulo

Atividade 2.6 (5 min)

2.6) Filtrar os resultados do exercício 4 não deixando aparecer as categorias Confections e Condiments

⋮

	CategoryName	Shipper...	subtotal ▾
1.	Meat/Poultry	2	1.354,81
2.	Meat/Poultry	3	1.002,51
3.	Produce	3	934,31
4.	Beverages	3	903,21
5.	Beverages	2	834,39
6.	Meat/Poultry	1	818,02
7.	Produce	2	812,03
8.	Dairy Products	1	782,22

1 - 18 / 18 < >

- Resolução

- Adicionar um filtro à tabela baseado, por exemplo, numa regex

Funções de Texto (algumas)

- **CONCAT(X, Y [, Z, ...])**
 - Retorna um texto que é a concatenação de X e Y (e Z e outros).
- **CONTAINS_TEXT(X, text)**
 - Retorna **TRUE** se X contiver texto. Caso contrário, retorna **FALSE**.
Diferencia maiúsculas de minúsculas.
- **REGEXP_EXTRACT(X, regular_expression)**
 - Retorna a primeira substring correspondente em "X", que corresponde ao padrão de expressão regular.

Atividade 2.7 (5 min)

2.7) Criar equipes/times dos vendedores (FirstName). Anne e Janet na Equipe 1, Michael e Steven na Equipe 2 e os outros na Equipe 3. Criar uma tabela mostrando o desempenho dessas 3 equipes em cada uma das categorias de produto.

	equipe	CategoryName	subtotal ▾
1.	Equipe3	Beverages	212.600,25
2.	Equipe3	Seafood	97.734,02
3.	Equipe3	Produce	76.786,5
4.	Equipe1	Beverages	70.153,25
5.	Equipe1	Seafood	38.161,31
6.	Equipe2	Beverages	26.828,75
7.	Equipe2	Produce	21.016,75
8.	Equipe1	Produce	13.591,75
9.	Equipe2	Seafood	13.164,2

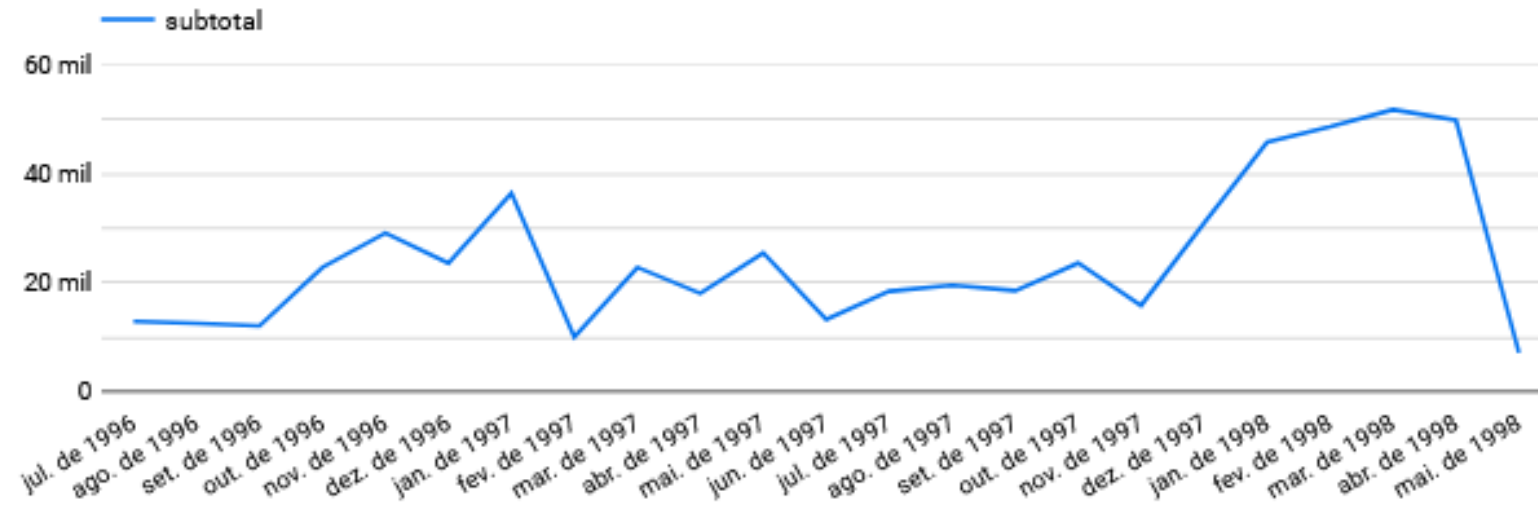
1 - 9 / 9 < >

- Resolução

- Criar um campo calculado para retornar o nome da equipe
- Adicionar as dimensões equipe e CategoryName
- Definir o campo da métrica (subtotal)
 - ✓ Assegurar-se de que a função de agregação é a que você deseja

Atividade 2.8 (5 min)

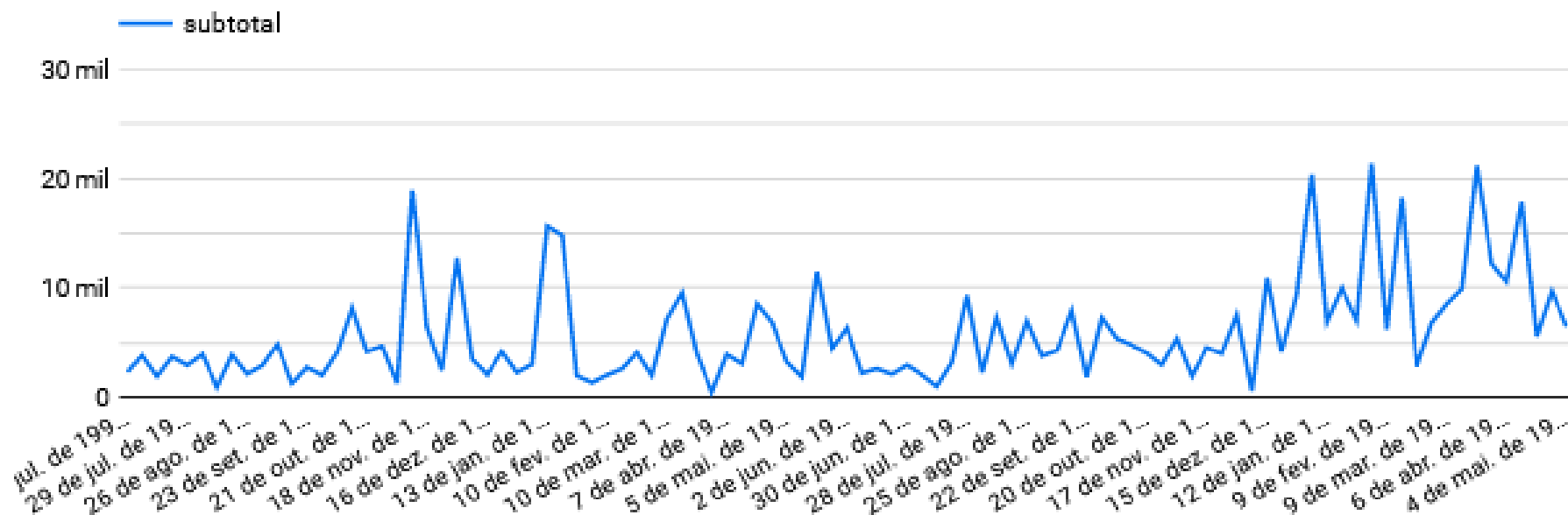
2.8) Criar um gráfico de linha do total de venda (\$) por mês



- Resolução
 - Altere o tipo da data no eixo X para mês e ano
 - Para visualizar a série acumulada, ative a respectiva opção na aba Estilo

Atividade 2.9 (5 min)

2.9) Criar um gráfico de linha do total de venda (\$) por semana



- Resolução

- Dica: altere o tipo da data no eixo X para semana e ano

Funções de Data

- CURRENT_DATE

- Retorna a **data** atual de acordo com o fuso horário especificado ou padrão.
- CURRENT_DATE("America/Sao_Paulo")
 - ✓ Fusos: https://en.wikipedia.org/wiki/List_of_tz_database_time_zones

- DATE_DIFF

- Diferença em número de dias

- DAY, HOUR, MONTH

Links úteis sobre funções

- Erros mais comuns ao usar funções
 - <https://www.optimizesmart.com/formula-rejection-in-google-data-studio/>
- Exemplos de uso das funções
 - <https://www.sumified.com/data-studio-case-function-examples/>

Criar um Scorecard

Produtos
29

Clientes
89

- Adicione um scorecard (visão geral) ao Dashboard
 - Inserir -> Visão geral (scorecard)
 - ✓ Arraste o campo calculado subtotal
 - para a região da métrica
 - Ou clique na métrica e defina o subtotal
 - Escolha a agregação apropriada
 - Defina o label (Nome)
- Algumas Funcionalidades
 - Definir um valor padrão
 - Agrupar Controles e Gráficos/Tabelas

Atividade 2.10 (5 min)

2.10) Criar um KPI (scorecard) com a quantidade total de produtos e outro com a quantidade total de clientes.



- Resolução
 - Adicionar um scorecard (visão geral) ao dashboard
 - Definir o campo da métrica

Recomendações de nomenclatura

- Use nomes consistentes (diminuem risco de equívocos)
 - Use prefixos de tipo
 - ✓ qt (quantidade), pct (%), dt (data)
 - E prefixos de temas
 - ✓ qtDistr qtVac
 - Exemplo: qtDistr1, qtDistr2, qtDistr3
 - qtVac1, qtVac2, qtVac3
 - Se estiver trabalhando em equipe,
 - ✓ Convenções são úteis para agilizar a comunicação
- Use uma regra de formação para nomes de filtros. Exemplo:
 - DATASOURCE_CAMPO_REGRA
 - ✓ IBGE_Populacao_ExcluirNA

Referências Bibliográficas

- KNAFLIC, C. N. (2018). Storytelling with data: a data visualization guide for business professionals.
- McKinney, W. (2018). Python for data analysis: Data wrangling with pandas, NumPy, and IPython.
- HURST, L. (2020). Hands on with Google Data Studio: a data citizen's survival guide.
<https://onlinelibrary.wiley.com/doi/book/10.1002/9781119616238>.
- WEXLER, S., SHAFFER, J., & COTGREAVE, A. (2017). The big book of dashboards: visualizing your data using real-world business scenarios.
- <https://www.storytellingwithdata.com/podcast>
- <https://seaborn.pydata.org/>
- <https://pandas.pydata.org/docs/>
- <https://d3js.org>
- <https://numpy.org/doc/stable/reference/index.html>