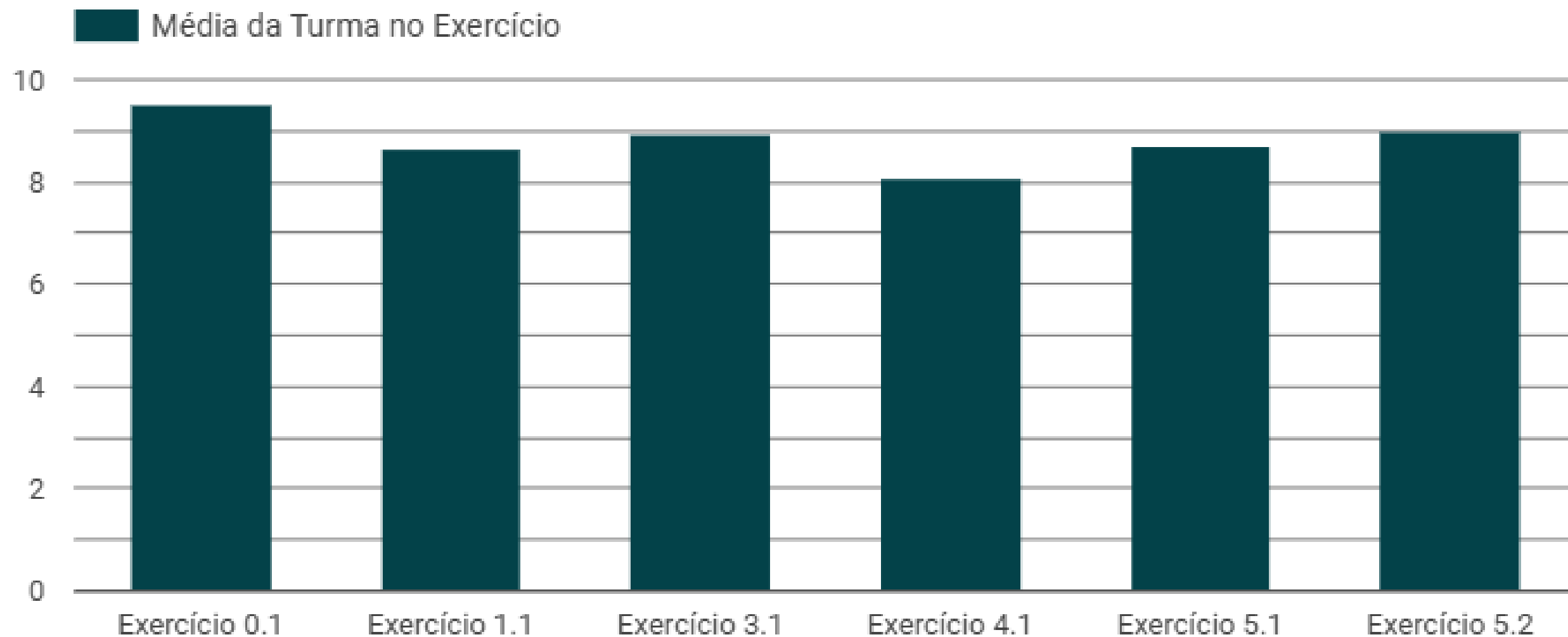


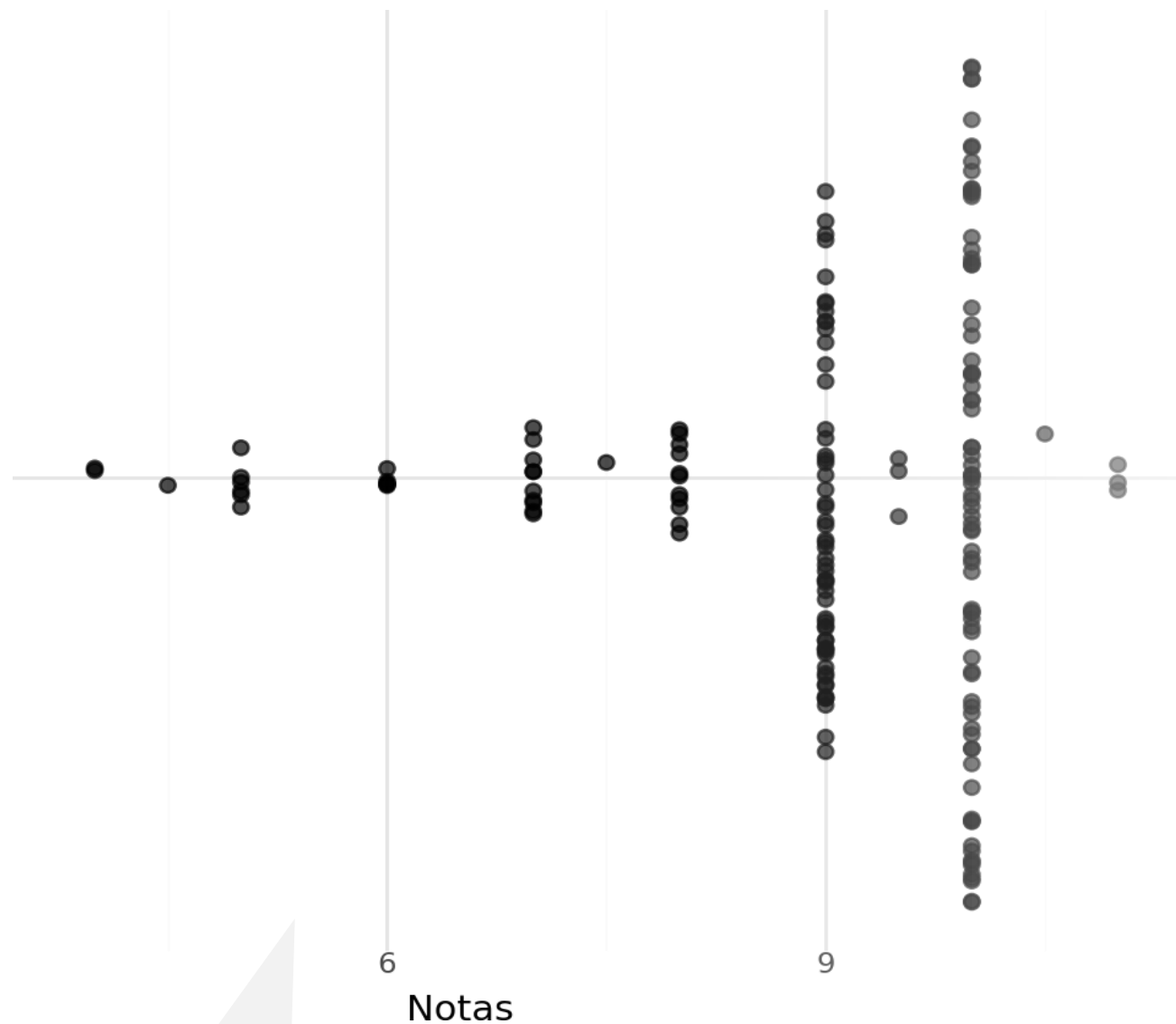
Introdução a Ciência de Dados



Professor: Alex Pereira

Desempenho da Turma





Feito no plotnine

Desempenho da Turma – Distribuição de Notas

Consequências para a
Administração Pública ?

Resolução do Exercício 3.1

- PIB Percapita da UF

- Média do PIB Percapita dos seus municípios



- ✓ Mentalizar os buckets (baldes) e as entidades (pontos) que estão em cada um
 - Numa base de dados pequena, é interessante carregar os dados brutos no painel e deixar por conta da ferramenta de BI calcular as métricas em diversos níveis de agregação

- SQL First

- Deslocar o processamento para o banco de dados

- ✓ Discutiremos os trade-offs no exercício 5.2

Resolução do Exercício 5.1 e 5.2

- Chave da tabela: id_municipio, ano, rede e **anos_escolares**
 - Desconto da nota para quem não notou este atributo na chave
 - ✓ Mentalizar os buckets (baldes) e as entidades (pontos) que estão em cada um
 - Não se importar por quais entidades/quantas estão no bucket e quem são os buckets
 - Foi o motivo do desconto na nota
- Solução despivote a coluna rede, e na sequência despivote a coluna anos_escolares
 - Ou filtre apenas um dos tipos de anos escolares (vide prox. Slide)
 - Dá pra fazer de outra maneira? Dá, eu pesquisei.



Query do exercício 5.1

```
SELECT * FROM
(
SELECT pibtab.*, ideb.rede, ideb.nota_saeb_media_padronizada
FROM `enapcd2021.pibpercapita` as pibtab
INNER JOIN `basedosdados.br_inep_ideb.municipio` as ideb
ON ideb.ano = pibtab.ano and ideb.id_municipio = pibtab.id_municipio
WHERE ideb.anos_escolares = 'iniciais (1-5)'
) as result
PIVOT (SUM(result.nota_saeb_media_padronizada) as nota_saeb FOR
result.rede in ('municipal', 'estadual', 'federal', 'publica'))
ORDER BY sigla_uf, id_municipio, ano
```

Habilidades e Habitos a serem adquiridos no contexto do avanço da IA

- Questionar o óbvio
 - Por que a multiplicação de dois números positivos resulta num número par e a multiplicação de dois números negativos também?
- Aprofundamento de entendimento/pesquisa
 - O porque em vários níveis
- Taste (gosto) / bom senso (arquitetura, estética, design)
- Dar nota/avaliar o trabalho de outro/IA
- Empatia (fonte: [Po-Shen Loh](#))



Growth mindset (Atitude de crescimento)

- Quando acreditamos que nossa
 - inteligência,
 - habilidades criativas e
 - caráter
 - ✓ são coisas que podemos melhorar significativamente.
 - Não saber é visto como uma oportunidade.



Por que (ou não) aprofundamos as pesquisas?

- Frequência com que você se depara com o problema
 - Fake until you make it ([Amy Cuddy](#))
 - ✓ the act of repeatedly behaving confidently can eventually help you internalize that confidence.
- Sensação de escassez
 - Livro [Scarcity](#) ([resumo](#) na perplexity)
 - ✓ a condição de ter **pouco tempo**, dinheiro ou outros recursos reconfigura a cognição, estreita o foco, **consome “largura de banda”** mental e cria armadilhas autorreforçadas
 - impulsiona comportamentos como tunneling, malabarismo e endividamento de curto prazo

Excelência baseada em intuição (Segundo Daniel Kahneman)

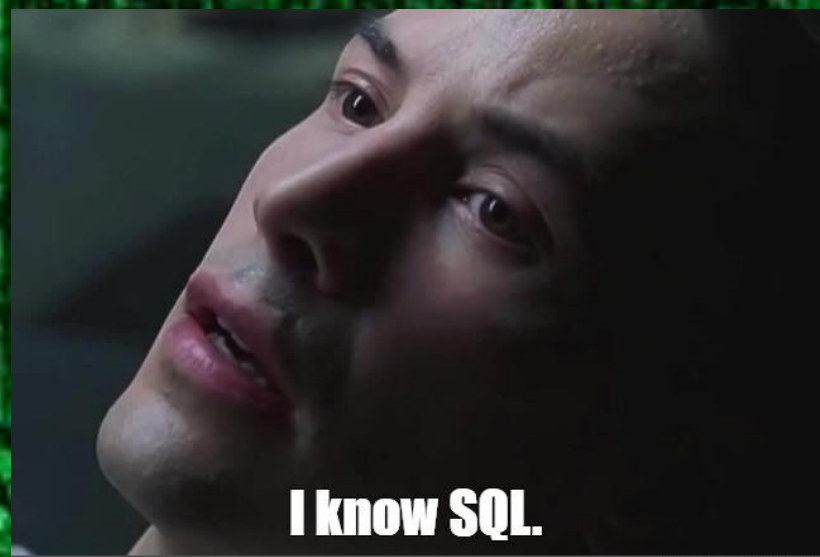
- Só emerge em “mundos de alta validade”, onde
 - as regularidades são estáveis,
 - há repetição abundante e
 - feedback rápido e inequívoco para calibrar o aprendizado
- Exemplos práticos
 - Domínios com excelência desenvolvível
 - ✓ xadrez, radiologia, arremesso esportivo, operações com protocolos estáveis.
 - Domínios com excelência intuitiva limitada
 - ✓ previsão macroeconômica, seleção de ações (stocks) no curto prazo

Trade-offs do Exercício 5.1 e 5.2

- Escala, performance e recursos
 - Volume/escala dos dados
 - Desempenho/latência
 - Uso de memória (RAM do notebook vs. DW)
 - Limites/cotas de plataforma
- Custo e movimentação de dados
 - Custo por bytes escaneados vs. compute local (**custo menor quando processado no bigquery**)
 - Tráfego de rede/egress (transferir dados para o Colab)
- Transformação e flexibilidade analítica
 - Complexidade/expressividade da transformação (joins, agregações, pivots, lógicas custom)
 - Flexibilidade e controle passo a passo (inspeção, merge, pivot_table, tratamento fino)
 - Confiabilidade em agregações/duplicatas (regras de aggfunc)
 - Qualidade/validação (anti-joins, contagens, asserts, checks de cardinalidade/nulos)

Trade-offs do Exercício 5.1 e 5.2

- Governança, segurança e padronização
 - Governança/auditoria/lineage e “**fonte única da verdade**”
 - Segurança/controle de acesso (IAM)
 - Padronização de schema/tipos/nomes
- Reprodutibilidade, operacionalização e integração
 - Reprodutibilidade e produção (incremental, agendamento, dbt/Airflow, scheduled queries)
 - Integração downstream (BI/Dashboards, consumo compartilhado)
 - Simplicidade/robustez operacional
- Manutenção, legibilidade e pessoas/objetivo
 - Legibilidade e manutenção do código/consulta
 - Perfil/skill da equipe (SQL vs. Python/pandas)
 - Objetivo/etapa do trabalho (**pipeline de dados vs. EDA/ML/visualização**)



I know SQL.

Possíveis aplicações de uma ferramenta de BI

- Construir um painel / dashboard
 - Painel de Compras do COVID-19
 - Painel sobre investimento em P&D
 - Painel de dados demográficos
 - Painel de Desempenho de website
- Contar uma história com dados
 - Em torno de um conjunto de visualizações
 - ✓ Menos personalizáveis para o usuário, do que as visualizações de painéis
 - História do Starbucks
 - Formula 1

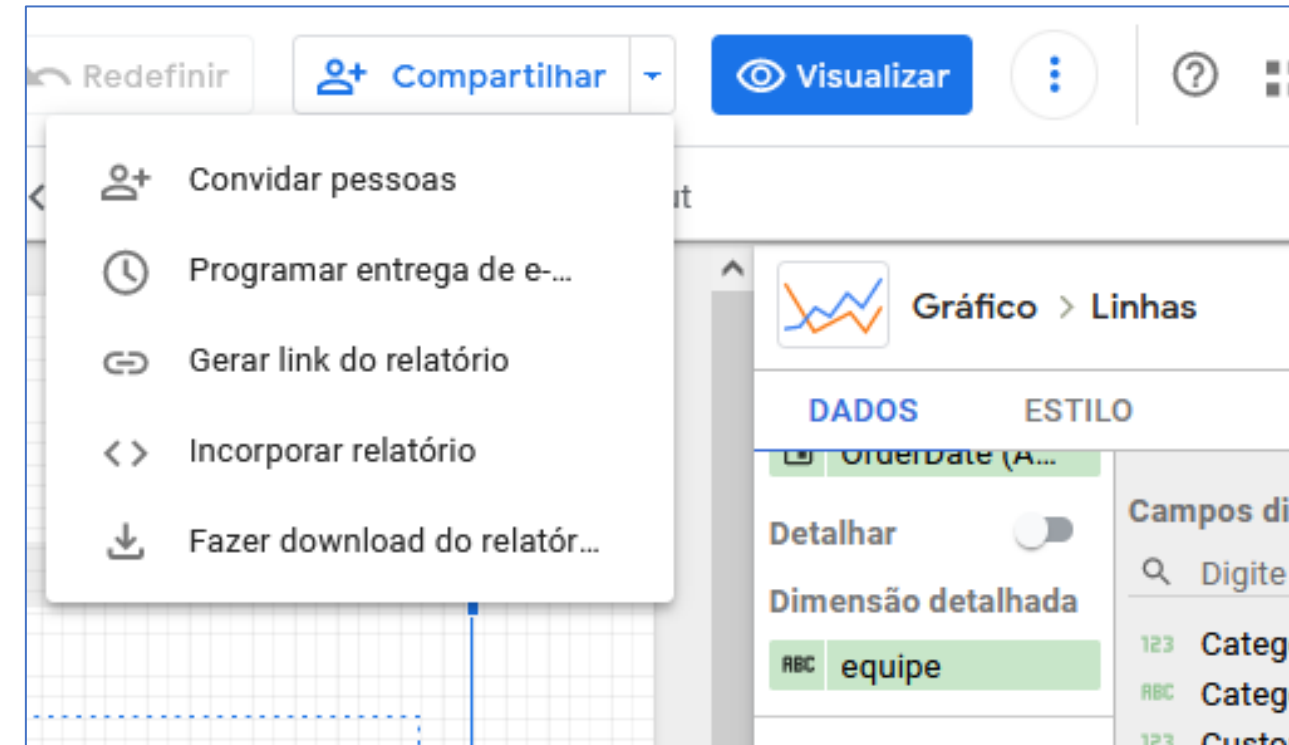
Google Looker Studio




<https://cloud.google.com/looker/docs/studio>

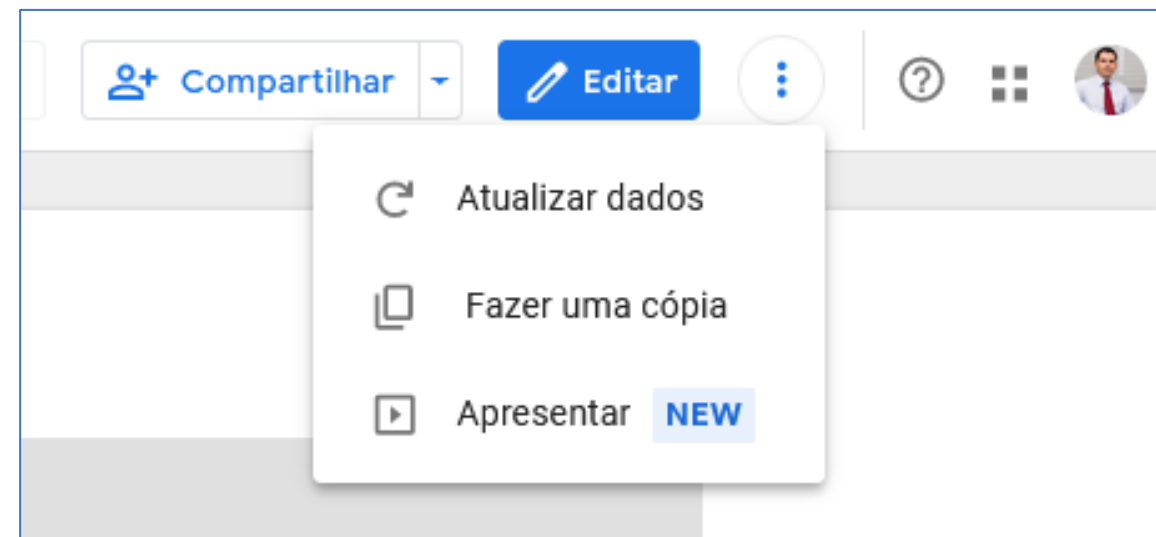
Conhecendo a interface do Google Looker Studio

- Compartilhar Relatório
- Programar entrega de e-mail
 - Pode-se criar um texto do e-mail
 - ✓ Envio de PDF
- Gerar link do relatório
- Incorporar relatório
 - Com IFRAME
- Fazer download do relatório
 - Em PDF



Conhecendo a interface do Google Looker Studio

- Três formas de editar um data source: [1](#), [2](#), [3](#)
 - [Deletar um campo calculado](#)
 - [Atualizar data source](#) com novos campos
 - [Compartilhar data source](#)
 - [Tornar data source acessível](#) a vários Relatórios/Painéis
- Configurações de páginas
 - [Nova Página](#), [Tamanho da Página](#), [Renomear](#), [Esconder no modo de visualização](#)
- 
 - Atualizar dados
 - Copiar relatório
 - Apresentar



Web Content Accessibility Guidelines 2.1 (WCAG)

- Um guia de boas práticas de acessibilidade
- 4 princípios de acessibilidade na Web
 - perceptível, operável, compreensível e robusto
- 3 níveis de Critérios de Sucesso
 - A (o mais baixo), AA e AAA (o mais elevado)
 - ✓ critérios objetivos e testáveis
 - permite que as WCAG 2.0 sejam utilizadas onde os requisitos e os testes de conformidade são necessários
 - ✓ tais como na especificação do projeto, nas compras, na regulamentação e nos acordos contratuais.
- Mínimo contraste
 - Nível AA: contraste de pelo menos 4.5:1 para texto normal e **3:1** para texto grande.
 - ✓ 3:1 para gráficos e componentes de interface do usuário (como bordas de entrada de formulário).
 - Nível AAA: **7:1** para texto normal e **4.5:1** para texto grande.
- Ferramenta Web para checar o contraste

Exemplos de Contraste

This is #000000 text on a #EA7439 background.

WCAG2 7.6:1 AAA

This is #FFFFFF text on a #EA7439 background.

WCAG2 1.7:1 ✗

This is #B5B7F6 text on a #494583 background.

WCAG2 4.5:1 AAA

This is #FFFFFF text on a #A088F1 background.

WCAG2 2.9:1 ✗

This is #027BB7 text on a #000000 background.

WCAG2 4.5:1 AAA

This is #4693CB text on a #FFFFFF background.

WCAG2 3.3:1 AA

Do Not Pass WCAG 2.0 AA
Normal Text Contrast Ratio



Buy Now



Buy Now



Buy Now

Do Pass WCAG 2.0 AA
Normal Text Contrast Ratio



Buy Now



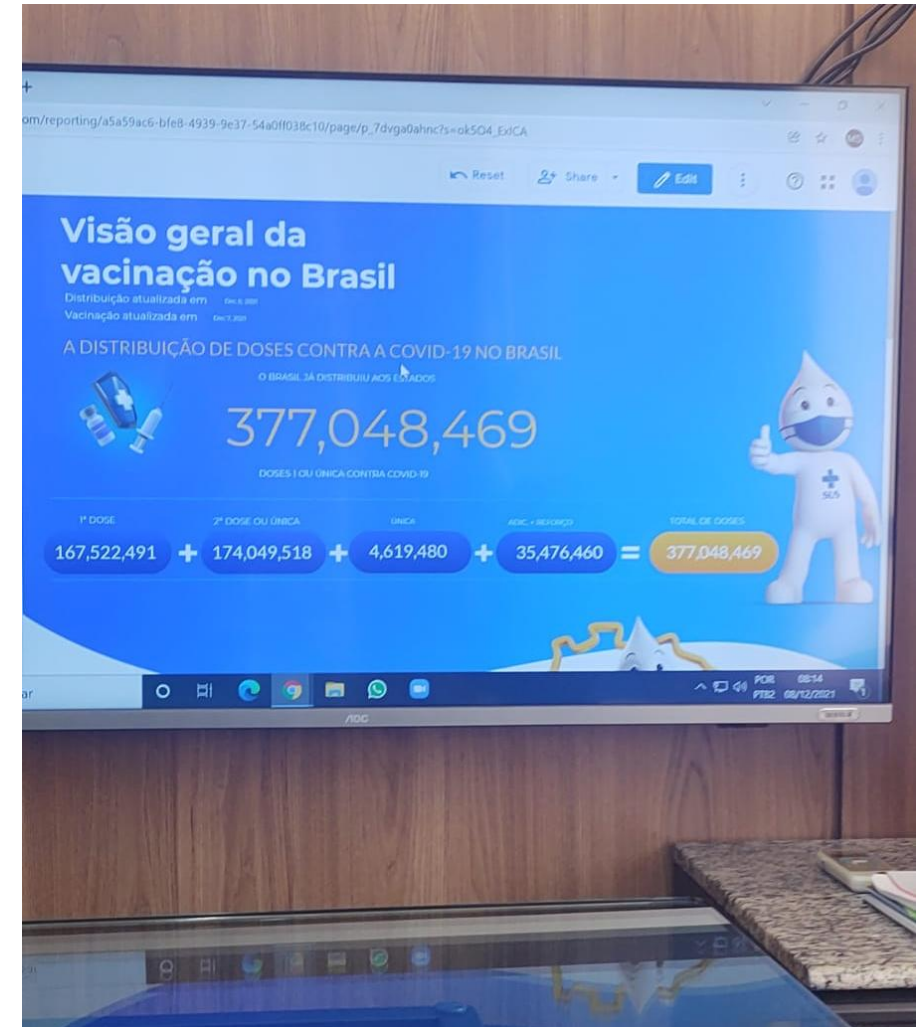
Buy Now



Buy Now

Conhecendo a interface do Google Looker Studio

- Personalizar tema
- Cor com Gradiente
- Formatação condicional em Tabela
- Nível de Relatório, Nível de Página
 - Report Level, Page Level
- Renomear rótulos (labels)
- Adicionar Imagem
 - Opção do menu Inserir
- Gerenciar Filtros
 - Opção do menu Recurso
- Limitar filtros dos controles
 - Basta agrupar o controle com os respectivos gráficos



Conhecendo a interface do Google Looker Studio

- Row Level Security

- Adicione um data source que contenha uma coluna com e-mails
 - ✓ Por exemplo, [desta planilha](#)
- [Ativar a opção "Filtrar dados por e-mail do visualizador"](#)
 - ✓ do respectivo data source
 - Substitua um dos emails pelo seu para testar

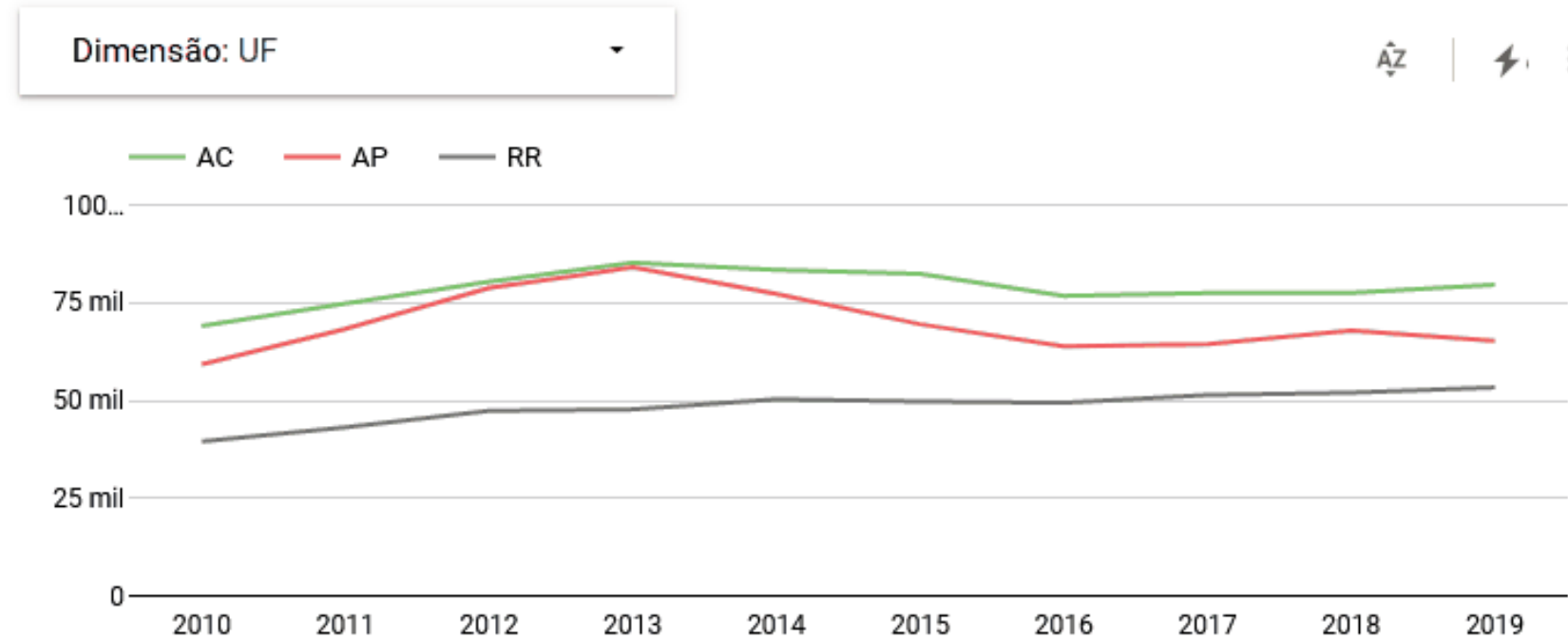
- [Serie acumulada](#)

- Histórico de Versões

- [Nomear versão atual](#)
- [Restaurar versão anterior](#)

Atividade 6.1 (5 min)

6.1) Mudar dimensão dinamicamente



- Demonstração

- Criar parâmetro para receber o valor da selecionado na lista
- Criar campo calculado para guardar o valor da dimensão detalhada
- Criar controle com lista suspensa
- Adicionar campo calculado ao gráfico no item dimensão detalhada

Atividade 6.1

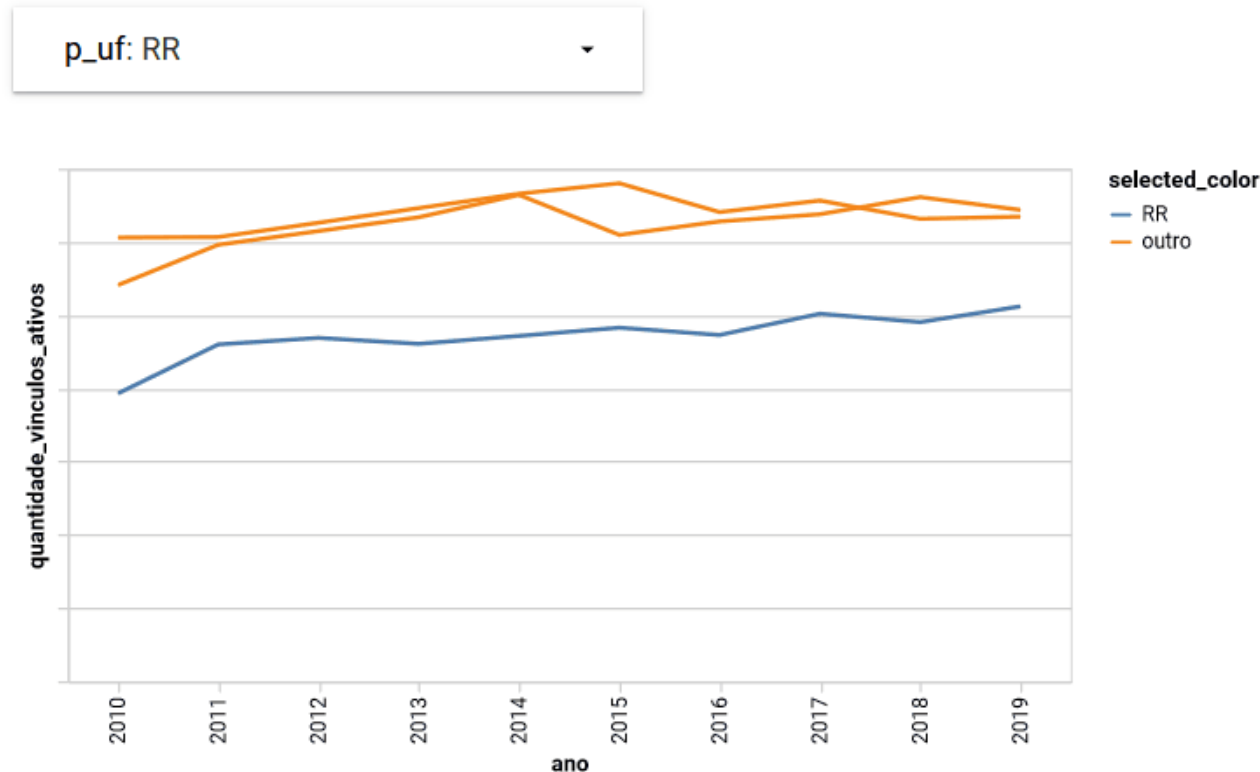
- Fórmula do campo calculado da dimensão detalhada

```
REGEXP_REPLACE(  
  REGEXP_REPLACE(  
    REGEXP_REPLACE(  
      CONCAT(p_dim, ";", sigla_uf, ";", cnae_1, ";", tipo_estabelecimento)  
      , "^(UF);(.*);(.*);(.*)", "\\2"  
    )  
    , "^(CNAE);(.*);(.*);(.*)", "\\3"  
  )  
  , "^(Tipo Estabelecimento);(.*);(.*);(.*)", "\\4"  
)
```

- REGEXP REPLACE(X, regular_expression, replacement)
 - argumento replacement: \\n é usado para substituir o conteúdo de X
 - ✓ pela substring do n-ésimo grupo
 - O que será retornado ?
 - ✓ REGEXP_REPLACE("UF;RO;522;CNPJ", "^(UF);(.*);(.*);(.*)", "\\2")
 - ✓ REGEXP_REPLACE("CNAE;RO;522;CNPJ ", "^(UF);(.*);(.*);(.*)", "\\2")

Atividade 6.2 (5 min)

6.2) Alterar a cor de uma das séries pela seleção de uma lista suspensa



- Demonstração

- Parâmetro, campo calculado e lista suspensa
- Gráfico de linha vega-lite (especificação no próximo slide)
 - ✓ vega-lite é uma biblioteca de gráficos interativos

Atividade 6.2 (5 min)

- Texto da configuração do gráfico de linha

```
{"$schema": "https://vega.github.io/schema/vega-lite/v4.0.2.json",  
  "mark": "line",  
  "encoding": {  
    "detail": {"type": "nominal", "field": "$dimension1"},  
    "color": {"type": "nominal", "field": "$dimension2"},  
    "y": {"type": "quantitative", "field": "$metric0", "axis": {"labels": false}},  
    "x": {"type": "nominal", "field": "$dimension0", "axis": {"labels": true}},  
    "tooltip": [  
      {"type": "nominal", "field": "$dimension0"},  
      {"type": "nominal", "field": "$dimension1"},  
      {"type": "quantitative", "field": "$metric0"}  
    ]  
  },  
  "height": 300, "width": 600  
}
```

- Fórmula do campo calculado:

```
case when REGEXP_MATCH(sigla_uf, p_uf) then p_uf else "outro" end
```

Centroide (Lat, Long) dos Municípios Brasileiros

- [Link para download](#) no site do IBGE
 - Link de [outra fonte](#)

The screenshot displays the IBGE website interface. On the left, a sidebar contains navigation links: 'O que é', 'Edições', 'Sobre a publicação', 'Acesso ao produto', and 'Informações técnicas'. The 'Edições' section is expanded, showing a list of editions. The edition '2010 - Cadastro de Localidades Seleccionadas' is highlighted with a red box. At the bottom of the sidebar, a dark blue button labeled 'Downloads' is also highlighted with a red box. The main content area is titled 'Downloads' and includes a note: 'Caso os arquivos não sejam exibidos, clique aqui para acessá-los através da página FTP.' Below this, a list of files is shown. The file 'Shapefile_SHP' is expanded, revealing a list of files: 'BR_Localidades_2010_v1.dbf', 'BR_Localidades_2010_v1.prj', 'BR_Localidades_2010_v1.shp', and 'BR_Localidades_2010_v1.shx'. The file 'BR_Localidades_2010_v1.dbf' is highlighted with a red box.

O que é

Edições

2010 - Cadastro de Localidades Seleccionadas

Sobre a publicação

Acesso ao produto

Informações técnicas

Downloads

Downloads

Caso os arquivos não sejam exibidos, clique aqui para acessá-los através da página FTP.

base_de_informacoes_indigenas_e_quilombolas_2019

cadastro_de_localidades_seleccionadas_2010

Geomedia_MDB

Google_KML

Shapefile_SHP

BR_Localidades_2010_v1.dbf

BR_Localidades_2010_v1.prj

BR_Localidades_2010_v1.shp

BR_Localidades_2010_v1.shx

Centroide (Lat, Long) dos Municípios Brasileiros

- Dado mantido pelo IBGE
 - Tabela de 21mil registros de localidades
 - ✓ Aldeia, aglomerado, vila, cidade, etc
 - contém
 - ✓ nome da localidade, categoria e subordinação político-administrativa, **coordenadas do centroide do setor censitário** de referência e altitude.

CD_GEO	NM_MUNICIP,C,60	NM_MICRO,C,100	NM_UF,C,60	CD_NIVEL	CD_CATEGOR,C,5	NM_CATEGOR,C,50	LONG,N,24	LAT,N,24,6	ALT,N,24,5	GMRotatio
1100015	ALTA FLORESTA D'OESTE	CACOAL	RONDÔNIA	1	05	CIDADE	-61,999824	-11,935540	337,73572	0,00000
1100015	ALTA FLORESTA D'OESTE	CACOAL	RONDÔNIA	2	15	VILA	-62,043898	-12,437239	215,24443	0,00000
1100015	ALTA FLORESTA D'OESTE	CACOAL	RONDÔNIA	2	20	VILA	-62,175549	-12,601415	181,04481	0,00000
1100015	ALTA FLORESTA D'OESTE	CACOAL	RONDÔNIA	2	25	VILA	-62,318650	-11,919792	191,57657	0,00000
1100015	ALTA FLORESTA D'OESTE	CACOAL	RONDÔNIA	2	30	VILA	-62,276812	-13,079806	157,28528	0,00000
1100015	ALTA FLORESTA D'OESTE	CACOAL	RONDÔNIA	2	35	VILA	-62,104428	-12,089439	407,70786	0,00000
1100023	ARIQUEMES	ARIQUEMES	RONDÔNIA	1	05	CIDADE	-63,033269	-9,908463	138,68898	0,00000
1100031	CABIXI	COLORADO DO OESTE	RONDÔNIA	1	05	CIDADE	-60,544314	-13,499763	236,06316	0,00000
1100031	CABIXI	COLORADO DO OESTE	RONDÔNIA	3	00001	POVOADO	-60,415206	-13,374447	264,99280	0,00000

Metadado da tabela de Localidades

Descrição dos Campos Finais para Features Geomedia, Shape e KML de Pontos de Localidades 2010 em 28/11/2011					
	Nome Campo Feature Geomedia e KML	Nome Campo Feature Shape	Tipo	Tamanho	Descrição
1	ID	ID	Autonumber	-	Contagem automática de geometrias ponto oriundas de setor
2	CD_GEOCODIGO	CD_GEOCODI	Text	20	Geocódigo do setor (15 dígitos numéricos)
3	TIPO	TIPO	Text	10	Classificação de Tipo (Urbano ou Rural, 6 dígitos alfa-numéricos)
4	CD_GEOCODBA	CD_GEOCODB	Text	20	Geocódigo do bairro (12 dígitos numéricos)
5	NM_BAIRRO	NM_BAIRRO	Text	60	Nome do bairro
6	CD_GEOCODSD	CD_GEOCODS	Text	20	Geocódigo do subdistrito (11 dígitos numéricos)
7	NM_SUBDISTRITO	NM_SUBDIST	Text	60	Nome do subdistrito
8	CD_GEOCODDS	CD_GEOCODD	Text	20	Geocódigo do distrito (9 dígitos numéricos)
9	NM_DISTRITO	NM_DISTRIT	Text	60	Nome do distrito
10	CD_GEOCODMU	CD_GEOCODM	Text	20	Geocódigo do Município (7 dígitos numéricos)
11	NM_MUNICIPIO	NM_MUNICIP	Text	60	Nome do Município
12	NM_MICRO	NM_MICRO	Text	100	Nome Micro-região
13	NM_MESO	NM_MESO	Text	100	Nome Meso-região
14	NM_UF	NM_UF	Text	60	Nome da UF
15	CD_NIVEL	CD_NIVEL	Text	1	Código do Nível da Localidade
16	CD_CATEGORIA	CD_CATEGOR	Text	5	Código da Categoria da Localidade
17	NM_CATEGORIA	NM_CATEGOR	Text	50	Nome da Categoria da Localidade
18	NM_LOCALIDADE	NM_LOCALID	Text	60	Nome da Localidade
19	LONG	LONG	Double	6 dec.	Longitude da Localidade em grau decimal
20	LAT	LAT	Double	6 dec.	Latitude da Localidade em grau decimal
21	ALT	ALT	Double	2 dec.	Altitude da Localidade, oriunda de SRTM em metros

Atividade 6.3 (10 min)

- Construir uma tabela com o centroide (Lat, Long) dos Municípios Brasileiros
 - e o código IBGE do respectivo município
 - ✓ Conforme o modelo a seguir:

	cod_ibge	categoria	long	lat	lat_long
0	1100015	CIDADE	-61.999824	-11.935540	-11.9355403048,-61.9998238963
6	1100023	CIDADE	-63.033269	-9.908463	-9.90846286657,-63.033269278
7	1100031	CIDADE	-60.544314	-13.499763	-13.4997634597,-60.5443135812
9	1100049	CIDADE	-61.442944	-11.433865	-11.4338650287,-61.4429442118
18	1100056	CIDADE	-60.818426	-13.195033	-13.195033032,-60.8184261647

Atividade 6.4 (5 min)

- Fazer um join da tabela de centroides dos municípios
 - com a tabela de PIB per capita do Exercício 3.1
 - ✓ A solução da atividade 6.3 encontra-se [aqui](#)
- Gravar o resultado no Bigquery
 - Com o mesmo nome da tabela que usou no exercício 3.1
 - ✓ Pois, aproveitaremos o data source no Looker Studio
- Atualize o data source do PIB per capita no Looker Studio
 - Para incorporar o novo campo lat_long ao modelo de dados

Atividade 6.5 (5 min)

- Trace um scatterplot (dispersão) no Looker Studio
 - O nome do [município na dimensão](#)
 - ✓ Na métrica X, a população e na Y o PIB dos municípios
 - Faça um [filtro para exibir pontos somente do ano de 2018](#)
 - [Use a escala logarítmica](#) (escala de registro)
 - ✓ Opção da aba Estilo
- Crie uma [lista suspensa](#)
 - e adicione o campo de UF (sigla_uf) a este controle
 - ✓ Defina o valor padrão como SP
- Se ainda não tiver resolvido o exercício 6.4
 - Use o código compartilhado [aqui](#)
 - ✓ para criar sua tabela com os dados dos municípios

Atividade 6.5 (5 min)

- Altere o valor inicial (Axis min)
 - dos dois eixos (X e Y)
 - ✓ X: 500 e Y: 10000000

Eixo Y

Show axis line ☒

Axis min

Axis max

Custom tick interval

Log scale ☒

Eixo X

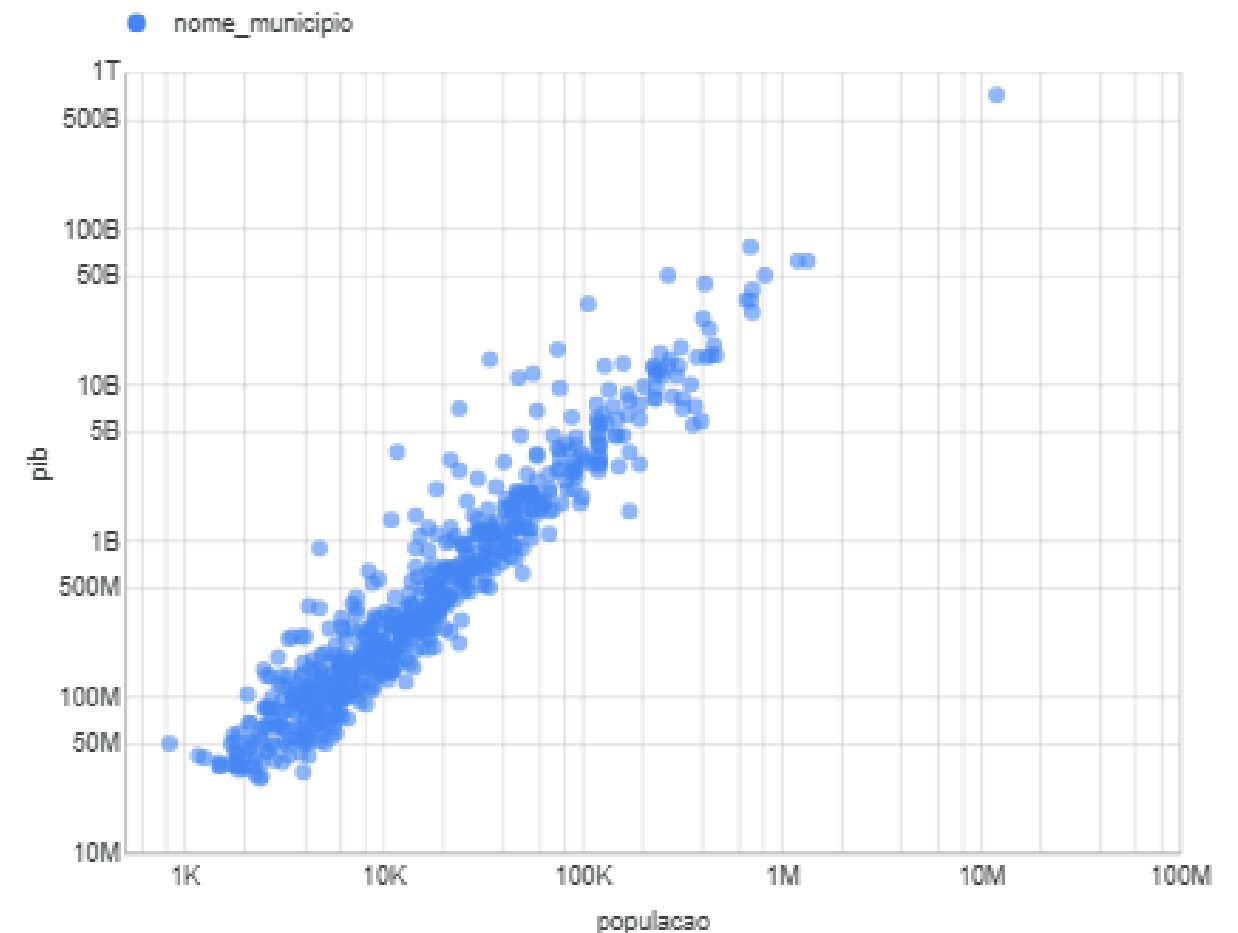
Show axis line ☒

Axis min

Axis max

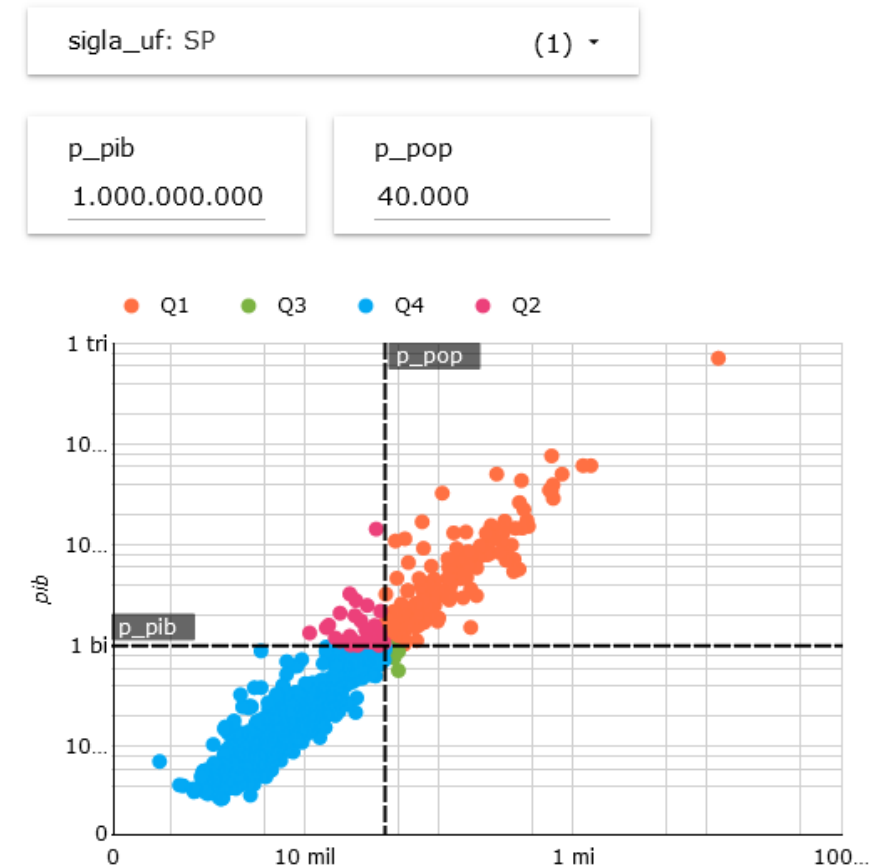
Custom tick interval

Log scale ☒



Atividade 6.6 (5 min)

- Crie os parâmetros p_pib e p_pop
 - com valores padrão 1 bilhão e 40 mil
- Crie duas Caixas de Entrada (controle)
 - E atribua a eles os parâmetros p_pib e p_pop
- Na aba Estilo, crie duas linhas de referência
 - do tipo Parâmetro
 - ✓ E atribua os parâmetros p_pib e p_pop
- Crie um campo calculado para definir as cores
 - E o adicione numa nova dimensão do scatterplot
 - ✓ Uma sugestão de fórmula encontra-se no próximo slide
- Defina cores para os quadrantes



Atividade 6.6 (5 min)

- Fórmula para o campo calculado das cores

case

when pib >= p_pib **and** populacao >= p_pop **then** "Q1"

when pib >= p_pib **and** populacao < p_pop **then** "Q2"

when pib < p_pib **and** populacao >= p_pop **then** "Q3"

else "Q4"

end

Atividade 6.7 (5 min)

- Converta o campo lat long para o tipo Informações Geográficas
 - e sub-tipo Latitude,Longitude
- Crie um gráfico do tipo Mapa de Balão
 - O atributo lat_long será atribuído automaticamente
 - ✓ para o campo Local
 - Perceba que este gráfico também reage ao filtro de UF
 - Atribua o campo calculado à Dimensão de Cor
 - Atribua o pib per capita ao Tamanho
- Ajuste detalhes na aba Estilo

Atividade 6.8 (até o final da aula)

- Crie uma query SQL para contabilizar
 - A quantidade de vacinas do covid, por dose (1ª, 2ª, Reforço e Adicional), UF, semana epidemiológica e tipo de imunizante
- Utilize a tabela de vacinação disponível no BigQuery
- Faça um filtro pelos estados do AC, AP e RR
 - Para limitar a quantidade de registros retornados
- Query inicial
 - ```
SELECT vacina,
sum(case when dose='1ª Dose' then 1 else 0 end) as qt_D1
FROM `basedosdados.br_ms_vacinacao_covid19.microdados_vacinacao`
where sigla_uf in ('AC','AP','RR')
group by vacina
limit 10;
```
  - ✓ Documentação da basedosdados.org

## Atividade 6.8 - Solução

```
SELECT sigla_uf, vacina
 ,(case
 when vacina='86' then 'Coronavac'
 when vacina='87' then 'Pfizer'
 when vacina='88' then 'Janssen'
 else 'Astrazeneca'
 end) as vacina_apelido
 ,EXTRACT(WEEK FROM data_aplicacao) AS semana
 ,EXTRACT(YEAR FROM data_aplicacao) AS ano
 ,EXTRACT(MONTH FROM data_aplicacao) as mes
 ,count(*) as qt_total
 ,sum(case when regexp_contains(dose,'1ª Dose$|^Dose$|Inicial') then 1 else 0 end) as qt_D1
 ,sum(case when regexp_contains(dose,'2ª Dose$') then 1 else 0 end) as qt_D2
 ,sum(case when regexp_contains(dose,'Reforço') then 1 else 0 end) as qt_Reforco
 ,sum(case when regexp_contains(dose,'Dose Adicional|3ª Dose') then 1 else 0 end) as qt_Adicional
 ,sum(case when regexp_contains(dose,'Única') then 1 else 0 end) as qt_Unica
FROM `basedosdados.br_ms_vacinacao_covid19.microdados_vacinacao`
where vacina in ('85','86','87','88','89')
group by sigla_uf, vacina, semana, mes, ano
```