

Introdução a Ciência de Dados



Professor: Alex Pereira

Possíveis aplicações de uma ferramenta de BI

- Construir um painel / dashboard
 - Painel de [Compras do COVID-19](#)
 - Painéis disponibilizam mais formas de exploração de dados
- Contar uma história com dados
 - [Poverty Climate Action](#)
 - ✓ [Vídeo](#) sobre esse dashboard
 - que ganhou uma competição de visualização de dados
 - Uma história tem uma narrativa
 - ✓ Em torno de um conjunto de visualizações
 - Menos personalizáveis para o usuário, do que as visualizações de painéis

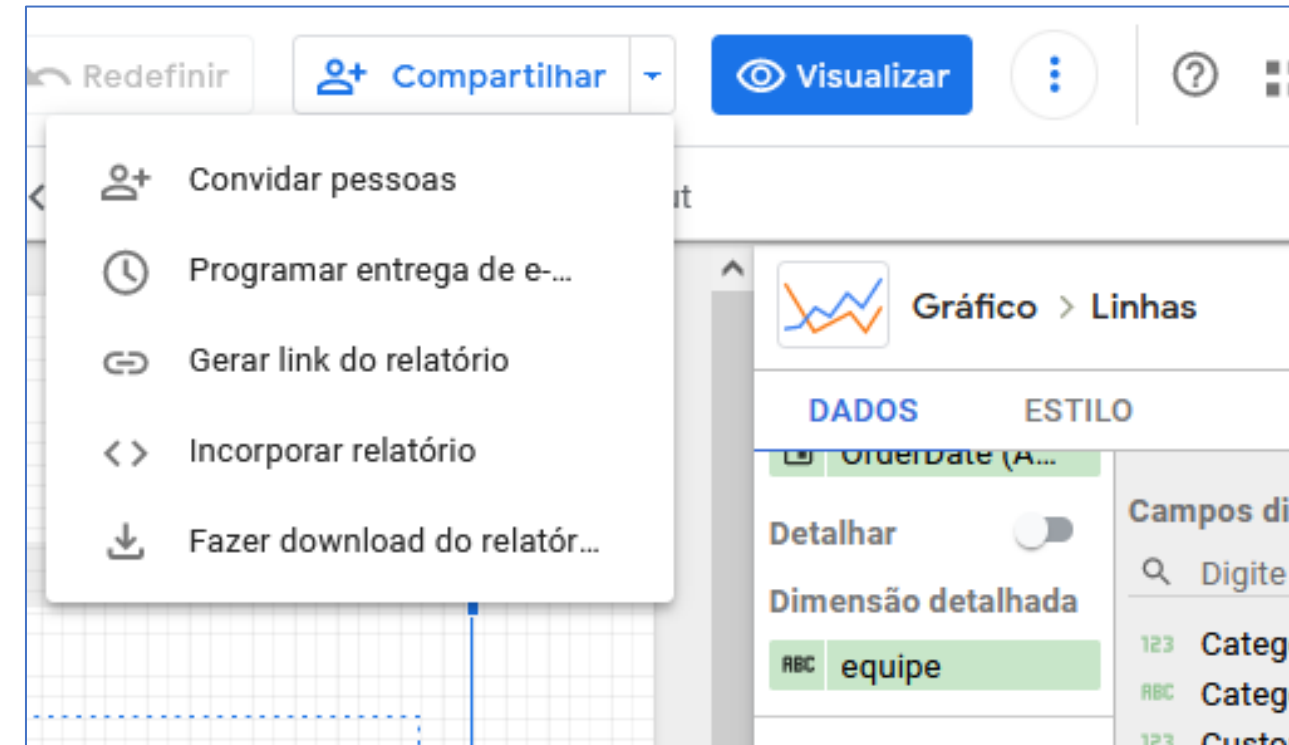
Google Data Studio



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


Conhecendo a interface do Google Data Studio

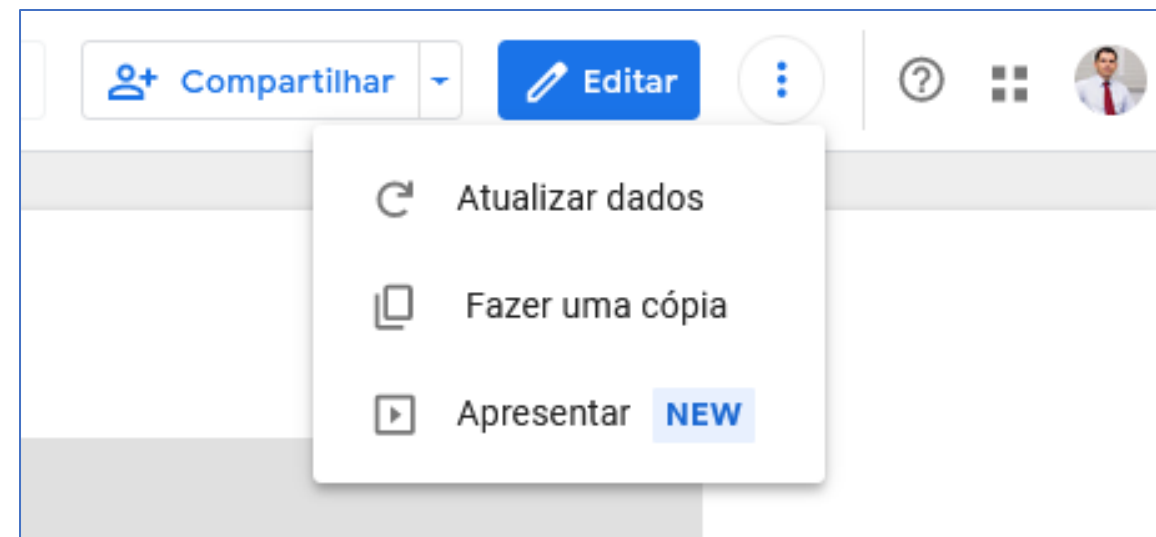
- Compartilhar Relatório
- Programar entrega de e-mail
 - Pode-se criar um texto do e-mail*
- Gerar link do relatório
- Incorporar relatório
- Fazer download do relatório
 - Com IFRAME



* [Script para envio de e-mails personalizados em massa](#)

Conhecendo a interface do Google Data 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 2.0 (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

Conhecendo a interface do Google Data 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 Data 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

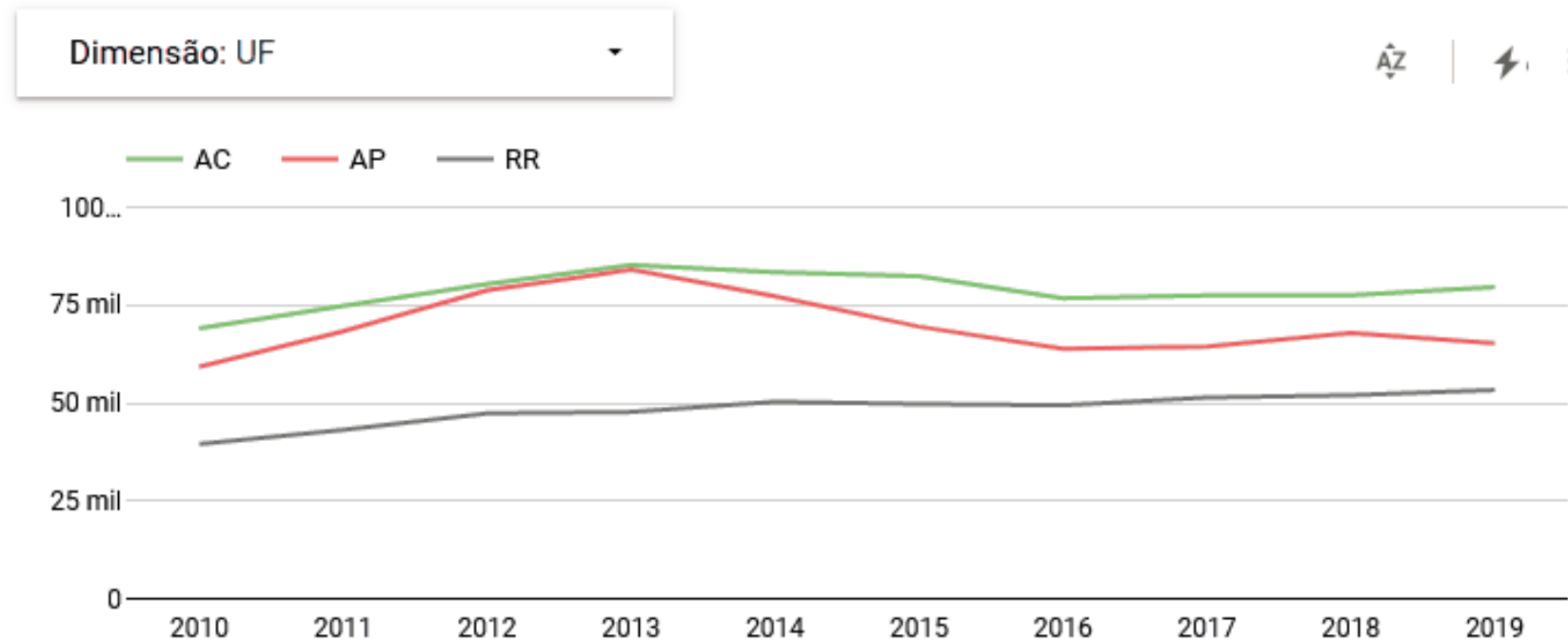
- [Serie acumulada](#)

- Histórico de Versões

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

Atividade 6.1 (5 min)

6.1) Mudar dimensão dinamicamente



- Demonstração

- [Criar um data source](#) da RAIS
 - ✓ `SELECT * FROM `enap-331414.enapdatasets.rais_AC_AP_RR_2010``
- [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

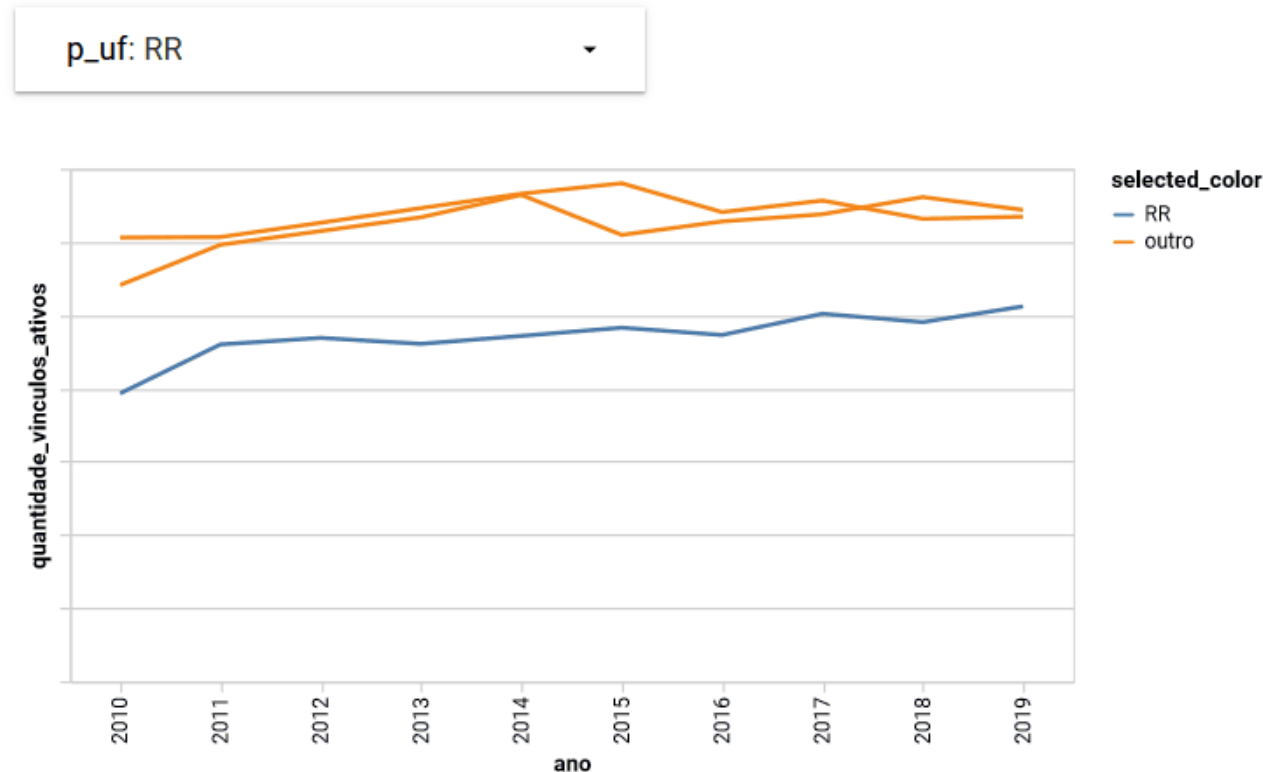
- Fórmula do campo calculado da dimensão detalhada

```
REGEXP_REPLACE(  
  REGEXP_REPLACE(  
    REGEXP_REPLACE(  
      CONCAT(p_dim, ";", sigla_uf, ";", cnae_1, ";", tipo_estabelecimento)  
      , "^(UF);(.*);(.*);(.*)", "\\2"  
    )  
    , "^(CNAE);(.*);(.*);(.*)", "\\3"  
  )  
  , "^(TipoEstabelecimento);(.*);(.*);(.*)", "\\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
 - ✓ 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

- Dado mantido pelo IBGE e disponível [aqui](#)
 - 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:

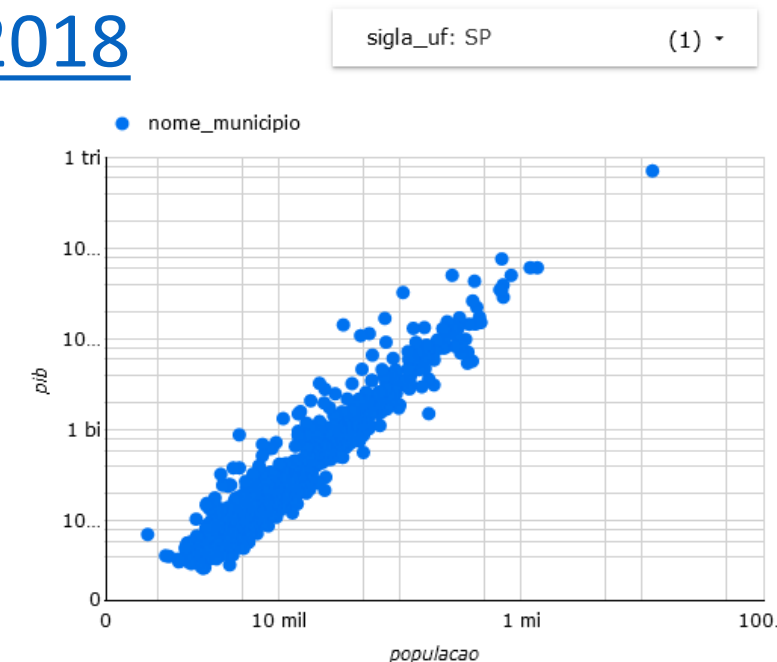
	cod_ibge	categoria	long	lat	lat_long
0	1100015	CIDADE	-61.999824	-11.935540	-11.9355403048,-61.9998238963
6	1100023	CIDADE	-63.033269	-9.908463	-9.90846286657,-63.033269278
7	1100031	CIDADE	-60.544314	-13.499763	-13.4997634597,-60.5443135812
9	1100049	CIDADE	-61.442944	-11.433865	-11.4338650287,-61.4429442118
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 Data Studio
- Atualize o data source do PIB per capita no Data Studio
 - Para incorporar o novo campo lat_long ao modelo de dados

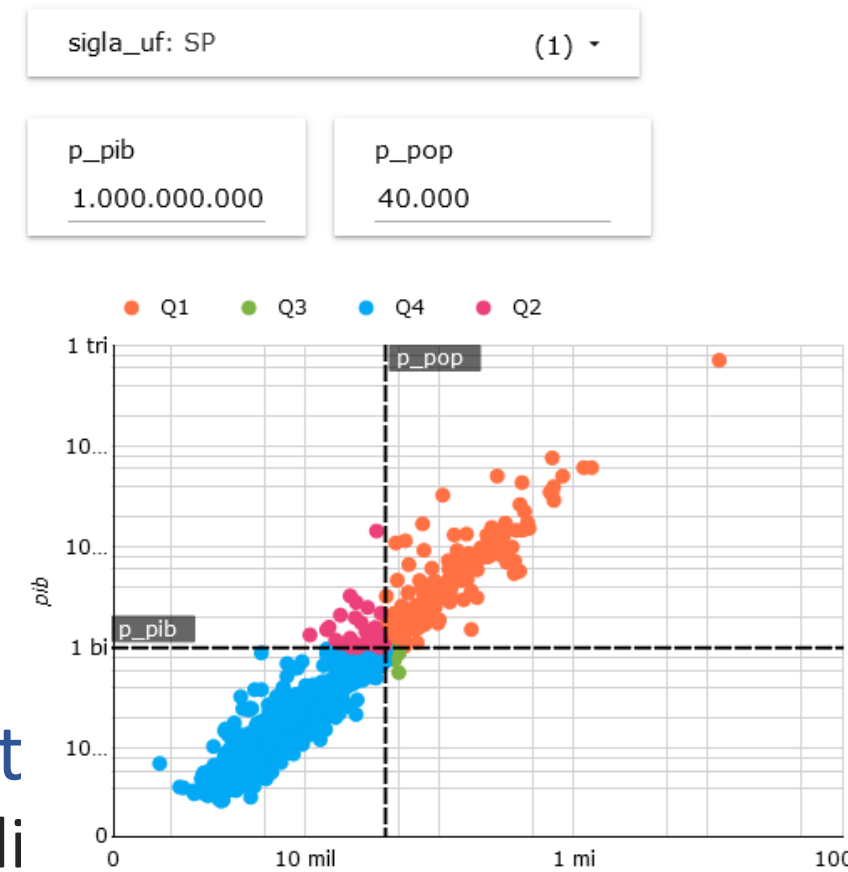
Atividade 6.5 (5 min)

- Trace um scatterplot (dispersão) no Data 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 6.6 (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 sli
- 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, por dose (1ª, 2ª, Reforço e Adicional), município, 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 * FROM
`basedosdados.br_ms_vacinacao_covid19.microdados_vacinacao`
LIMIT 100`
 - ✓ [Documentação](#) da basedosdados.org