FGV DIREITO RIO

Scatterplot e regplot: relações entre variáveis numéricas

Programação para Advogados – 2024.2

José Luiz Nunes e Lucas Thevenard

O que aprendemos até aqui?

- **Argumentos no seaborn**
  - `x` e `y`
  - `data`
  - `hue`
  - `color`
  - `palette`
  - `hue_order`
  - `errorbar`
  - `estimator`

O que aprendemos até aqui?

- **O que passamos para os argumentos?**
  - `x` e `y`:
    - variáveis numéricas ou categóricas - `str`
  - `data`:
    - nosso conjunto de dados - `pd.DataFrame`
  - `hue`:
    - variável categórica - `str`
  - `color`:
    - cor única - `str`

O que aprendemos até aqui?

- **O que passamos para os argumentos?**
  - `palette`:
    - nome de paleta ou conjunto de cores - `str`, `list`, ou `dict`
  - `hue_order`:
    - ordem das categorias - `list` de `str`
  - `errorbar`:
    - controla barra de erro - `None` (**não vimos como alterar**)
  - `estimator`:
    - função para resumir dados - `str`; `"mean"`, `"sum"`, `"median"`

Roteiro de Aula

- Scatterplot
- Regplot
    - O que é um modelo?
- Anotações em gráficos

Passos Preliminares

```python
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

idh_2022 = pd.read_csv("https://bit.ly/idh_tidy_2022")

idh_2022.head()
```

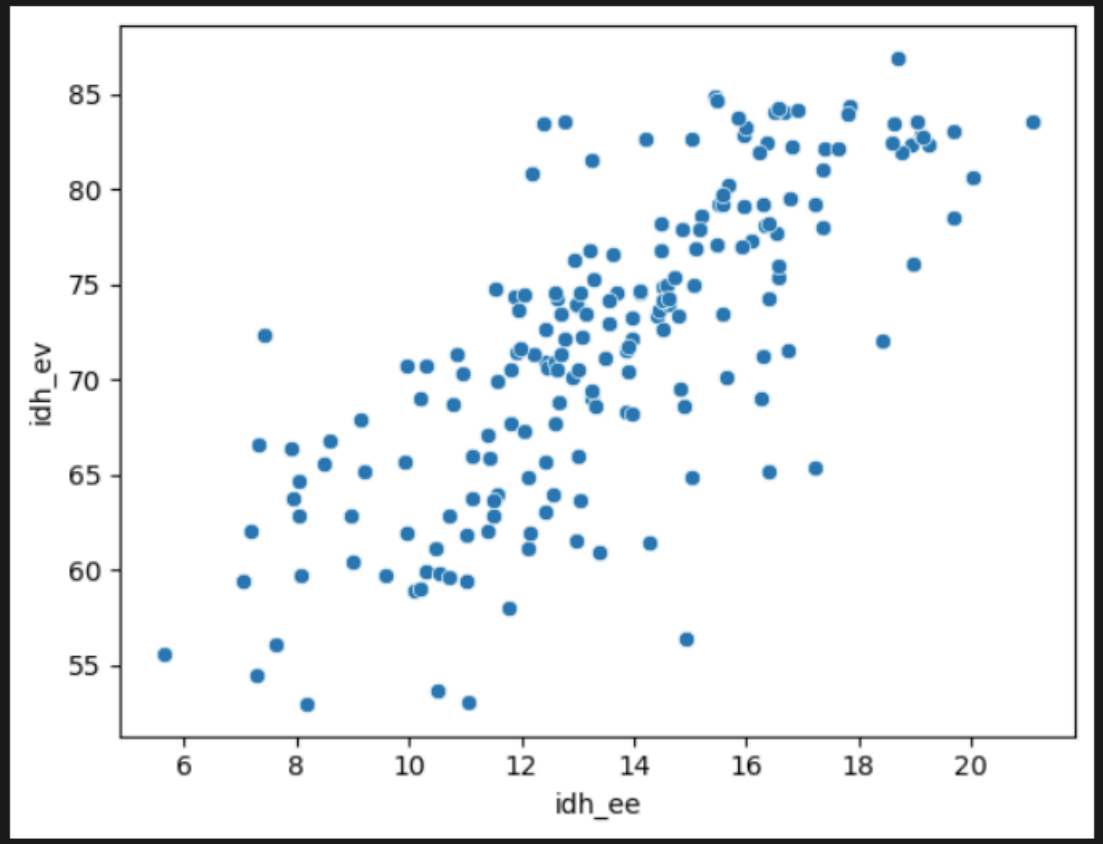
✓ 2.1s Python

	sigla	pais	grupo_idh	regiao	ranking_idh	idh	idh_ev	idh_ee	idh_me	
0	AFG	Afeganistão	Baixo	Ásia do Sul	182.0	0.462	62.879	10.705385	2.514790	1335
1	ALB	Albânia	Alto	Europa e Ásia Central	74.0	0.789	76.833	14.487470	10.121144	15293
2	DZA	Argélia	Alto	Países Árabes	93.0	0.745	77.129	15.487880	6.987444	10978
3	AND	Andorra	Muito Alto	NaN	35.0	0.884	83.552	12.783780	11.613440	54233.
4	AGO	Angola	Mediano	África Sub-sahariana	150.0	0.591	61.929	12.167600	5.844292	5327

5 rows × 27 columns

Scatterplot e regplot: relações entre variáveis numéricas

```
[10]  sns.scatterplot(x="idh_ee", y="idh_ev", data=idh_2022)                Python

...   <Axes: xlabel='idh_ee', ylabel='idh_ev'>
```

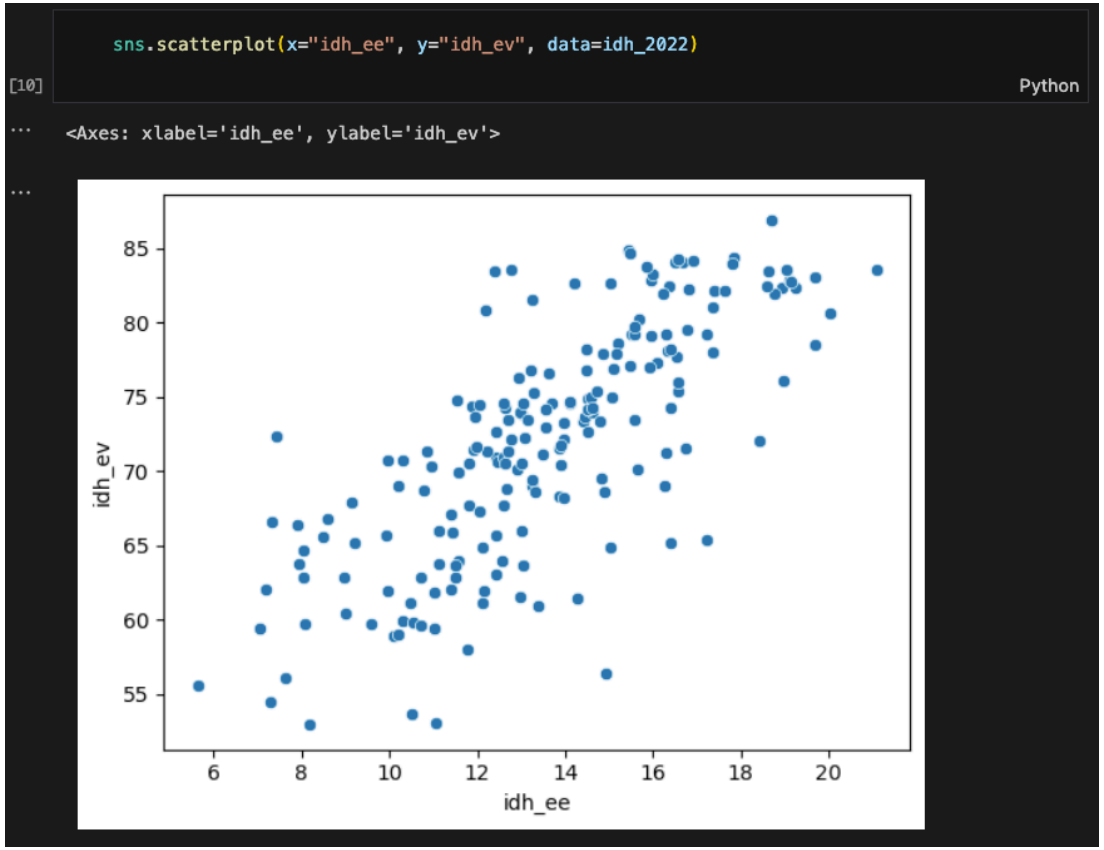


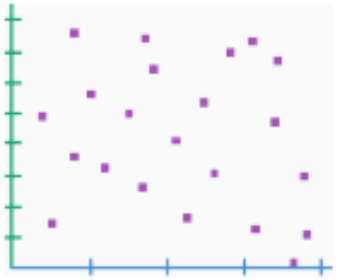
Convenção dos eixos

- Usualmente colocamos a variável que queremos entender/prever/explicar no **eixo y**.
  - Chamamos ela de "**variável de resposta**" ou "**variável dependente**".
- A variável utilizada para explicar o fenômeno é usualmente colocada no **eixo x**.
  - Chamamos ela de "**variável explicativa**" ou "**variável independente**".

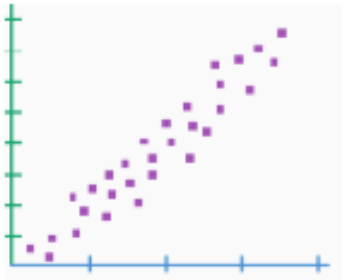
Scatterplot

- O gráfico ao lado é um scatterplot que relaciona a expectativa de vida dos países (eixo y) à expectativa de escolaridade (eixo x).
    - Qual é a relação entre essas variáveis evidenciada pelo gráfico mostra?

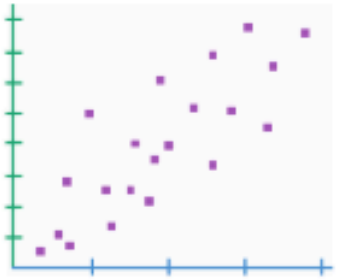




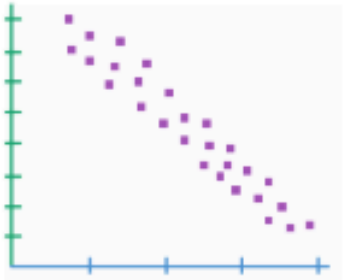
Sem correlação



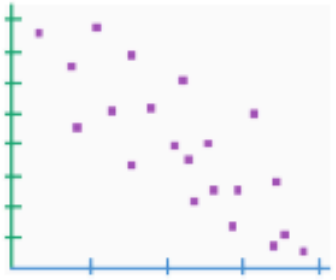
Correlação positiva forte



Correlação positiva média



Correlação negativa forte



Correlação negativa média

Cuidado: correlação ≠ causalidade

- O gráfico parece suportar a ideia de que quando aumentamos a expectativa de escolaridade de um país a expectativa de vida aumenta daquela população aumenta.
- No entanto essa conclusão é precipitada!
  - Quais outras explicações podemos ter?
    - **Hipótese rival 1**: A direção da causalidade é a inversa: é o aumento da expectativa de vida que causa aumento da escolaridade.
    - **Hipótese rival 2**: Há uma terceira variável que está causando tanto o aumento da expectativa de vida como de escolaridade.
    - **Hipótese rival 3**: Não há nenhuma relação real entre esas variáveis. A correlação observada é fruto do acaso (correlação espúria).

spurious correlations

correlation is not causation

random · discover · *next page* →

don't miss spurious scholar,
where each of these is an academic paper

Google searches for 'batman'

correlates with

The number of security guards in Oklahoma

Relative volume of Google searches for 'batman' (Worldwide, without quotes) · Source: Google Trends

BLS estimate of security guards in Oklahoma · Source: Bureau of Larbor Statistics

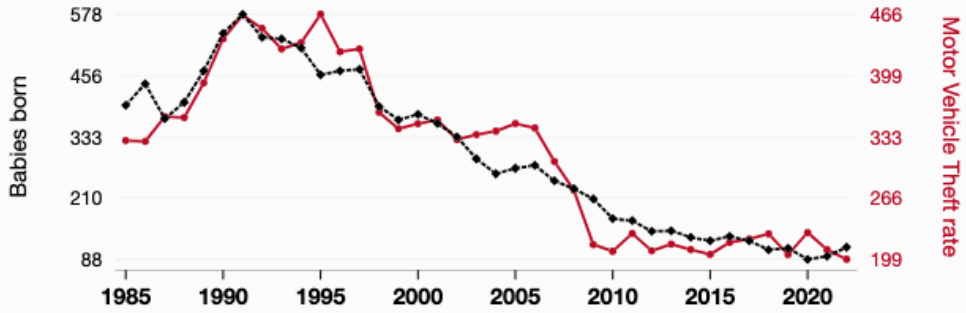
2004-2022, r=0.848, r²=0.719, p<0.01 · tylervigen.com/spurious/correlation/5227

www.tylervigen.com

Popularity of the first name Margarita

correlates with

Motor vehicle thefts in Indiana



Babies of all sexes born in the US named Margarita · Source: US Social Security Administration

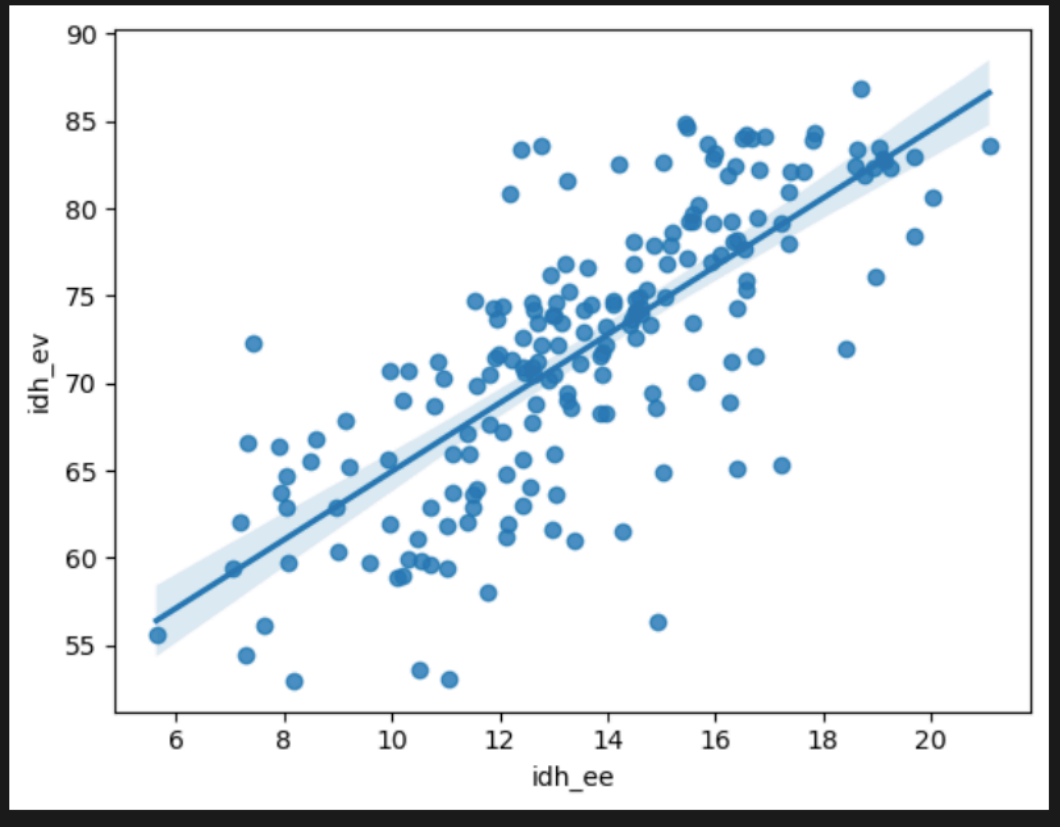
The motor vehicle theft rate per 100,000 residents in Indiana · Source: FBI Criminal Justice Information Services

1985-2022, r=0.956, r²=0.914, p<0.01 · tylervigen.com/spurious/correlation/2794

View details about correlation #2,794

Scatterplot e regplot: relações entre variáveis numéricas

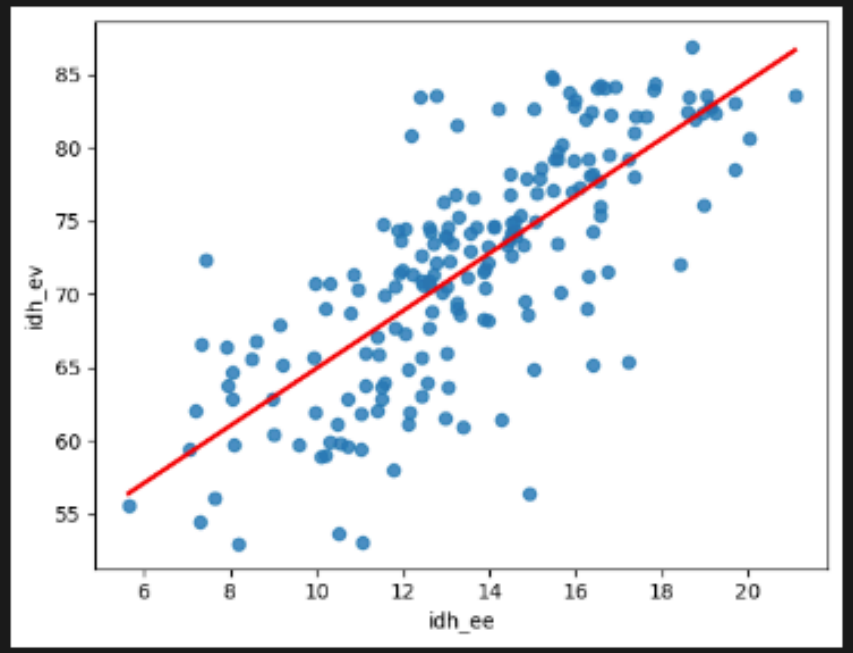
```
[12]  sns.regplot(x="idh_ee", y="idh_ev", data=idh_2022)                Python

...   <Axes: xlabel='idh_ee', ylabel='idh_ev'>
```



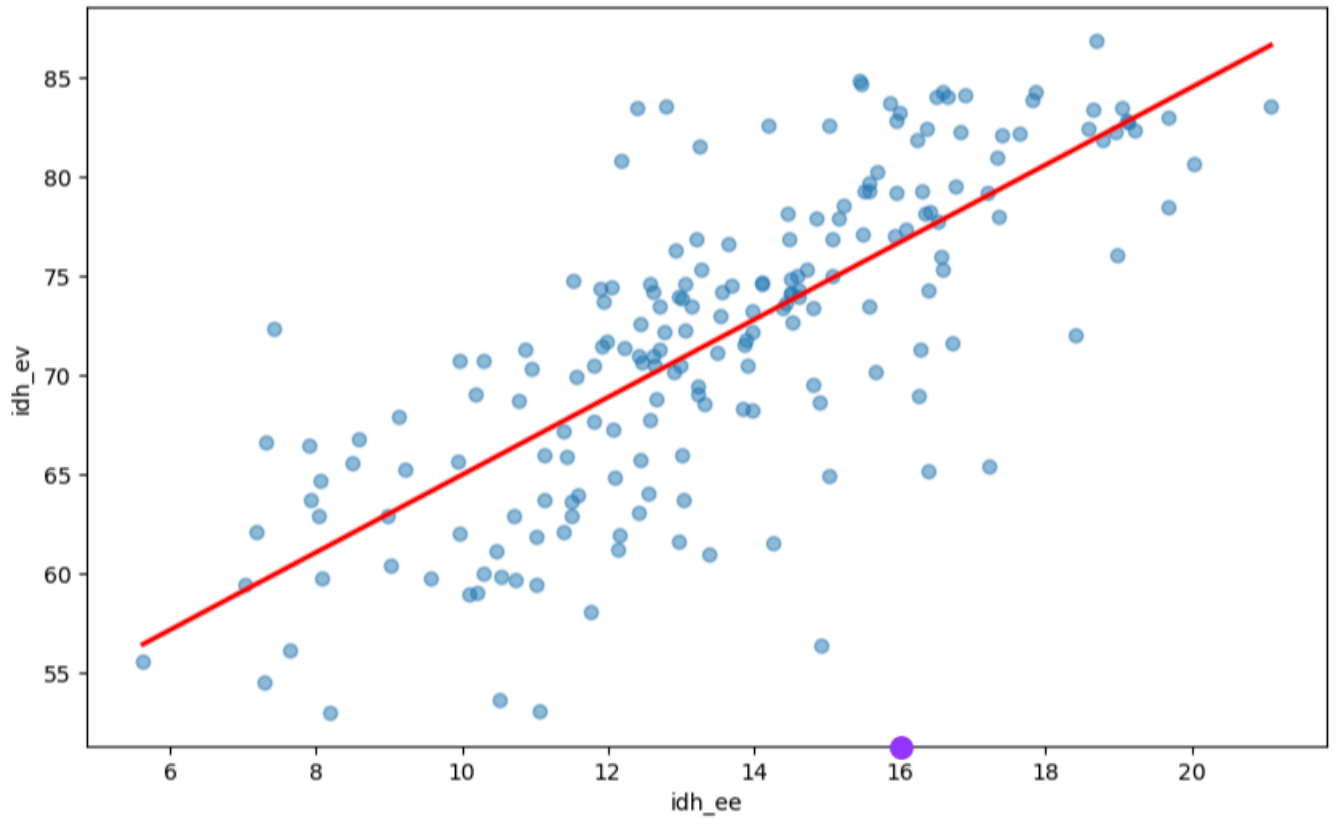
```python
sns.regplot(
    x="idh_ee",
    y="idh_ev",
    ci=None,
    line_kws={"color": "red"},
    data=idh_2022
)
```

[13] Python

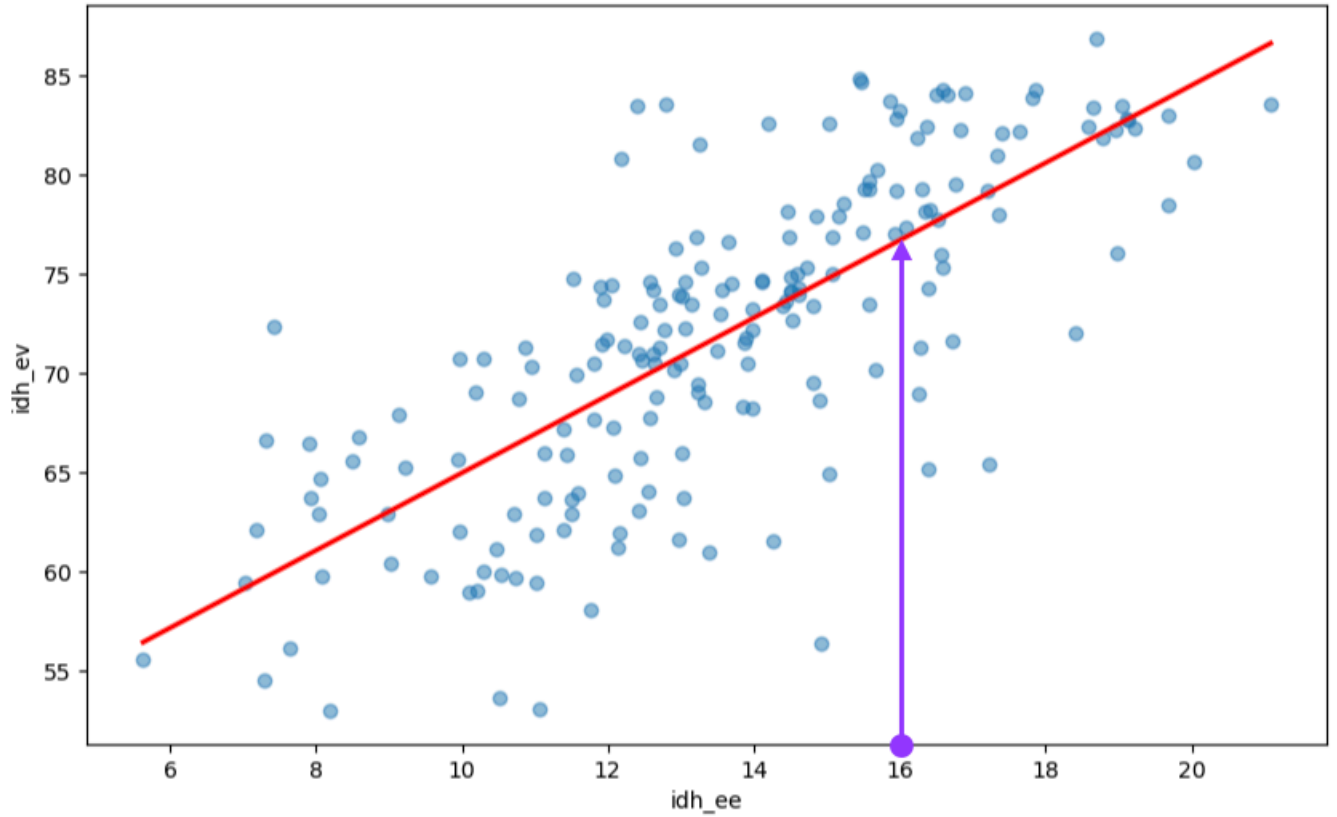
...



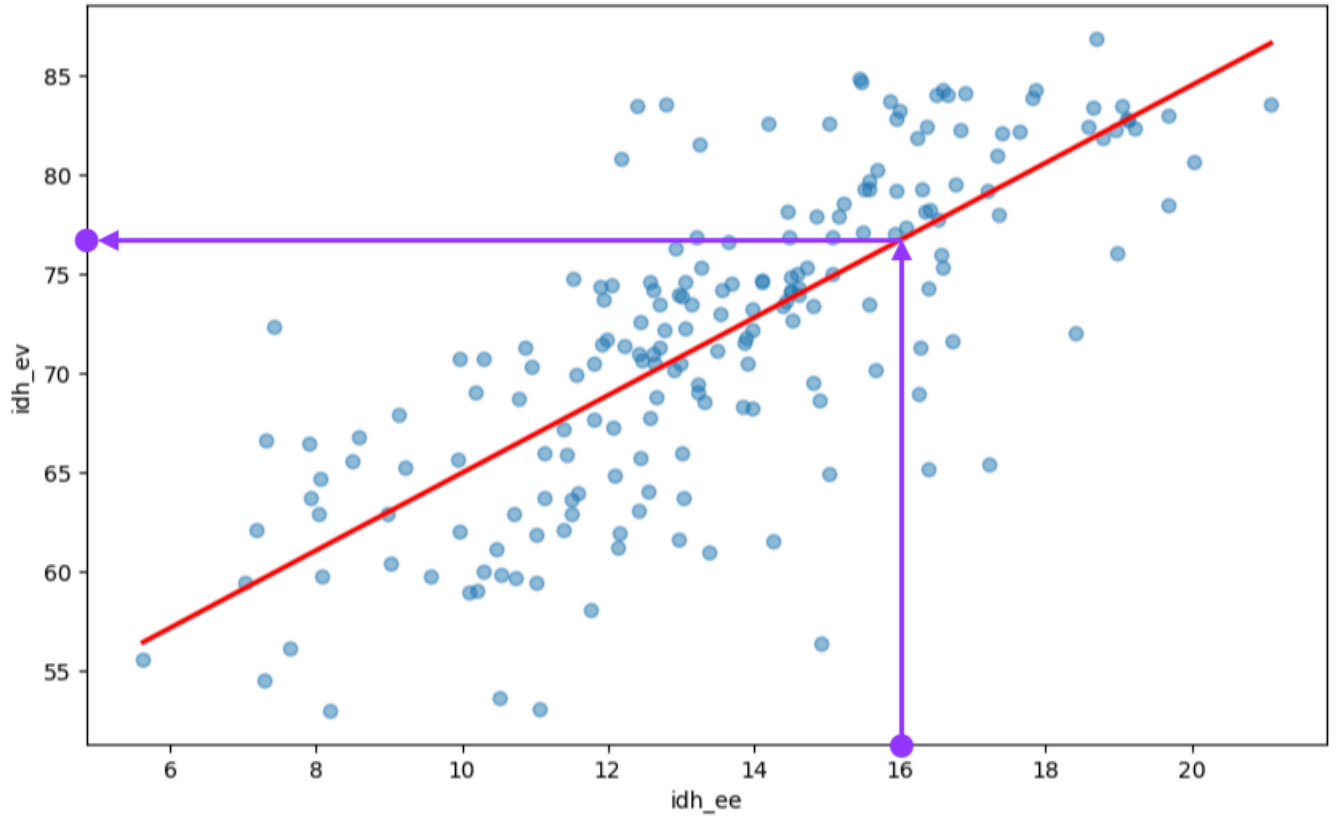
A reta como um modelo



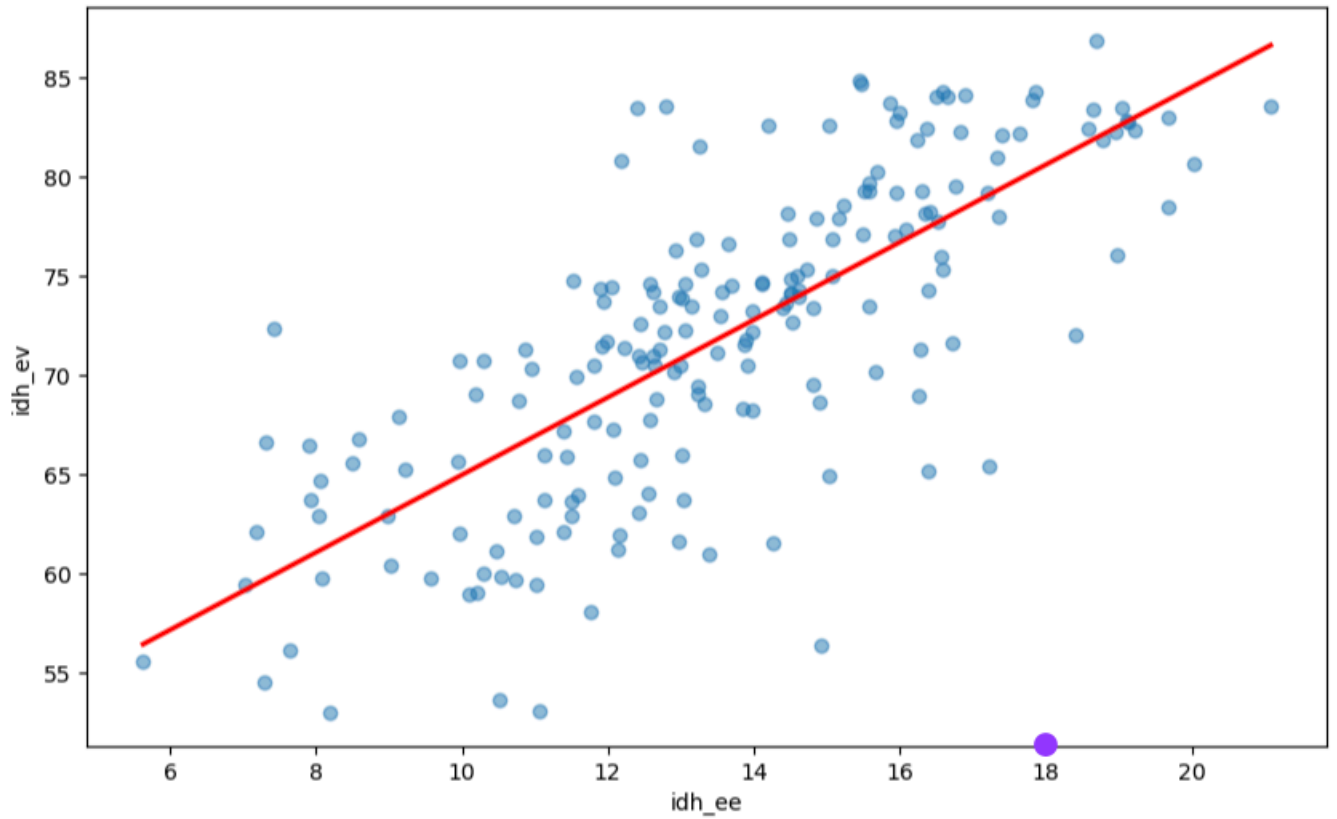
A reta como um modelo



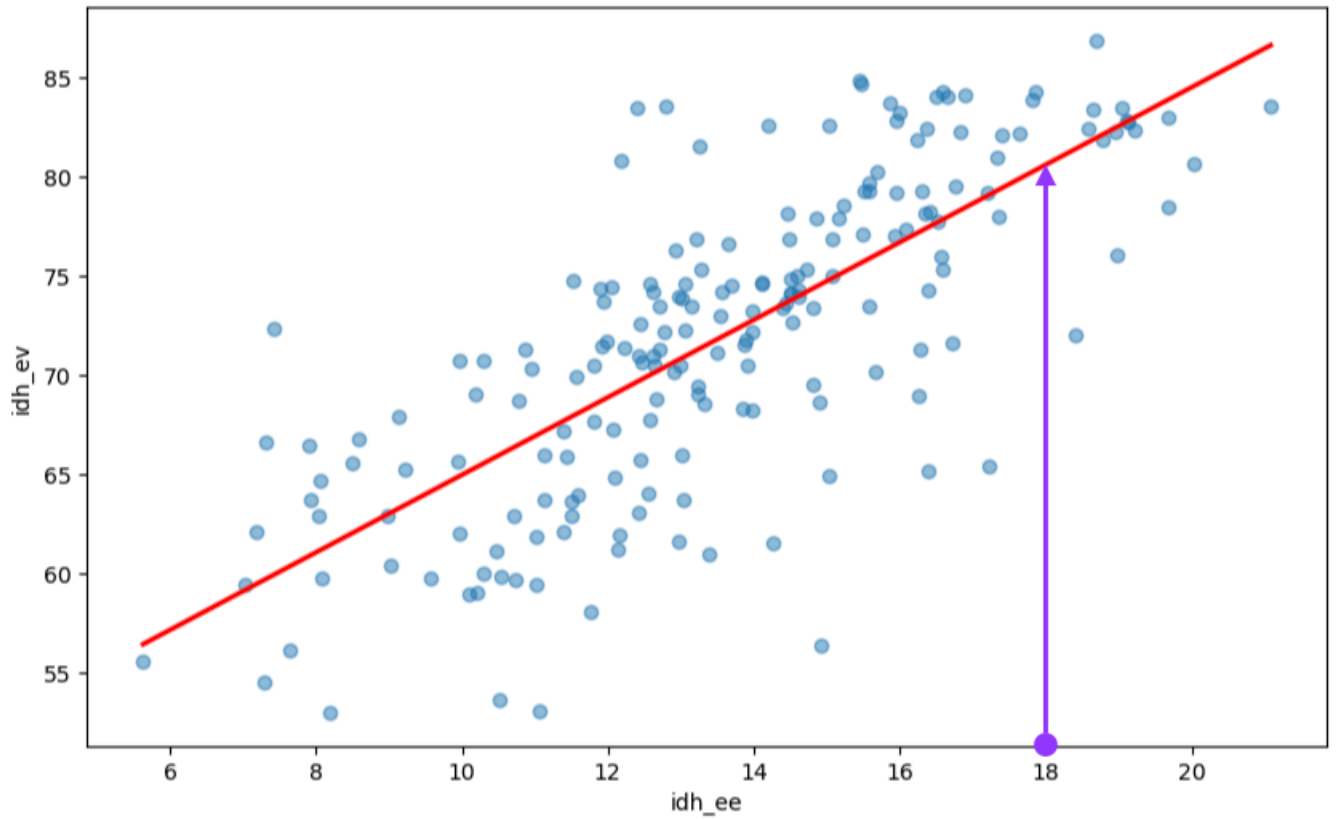
A reta como um modelo



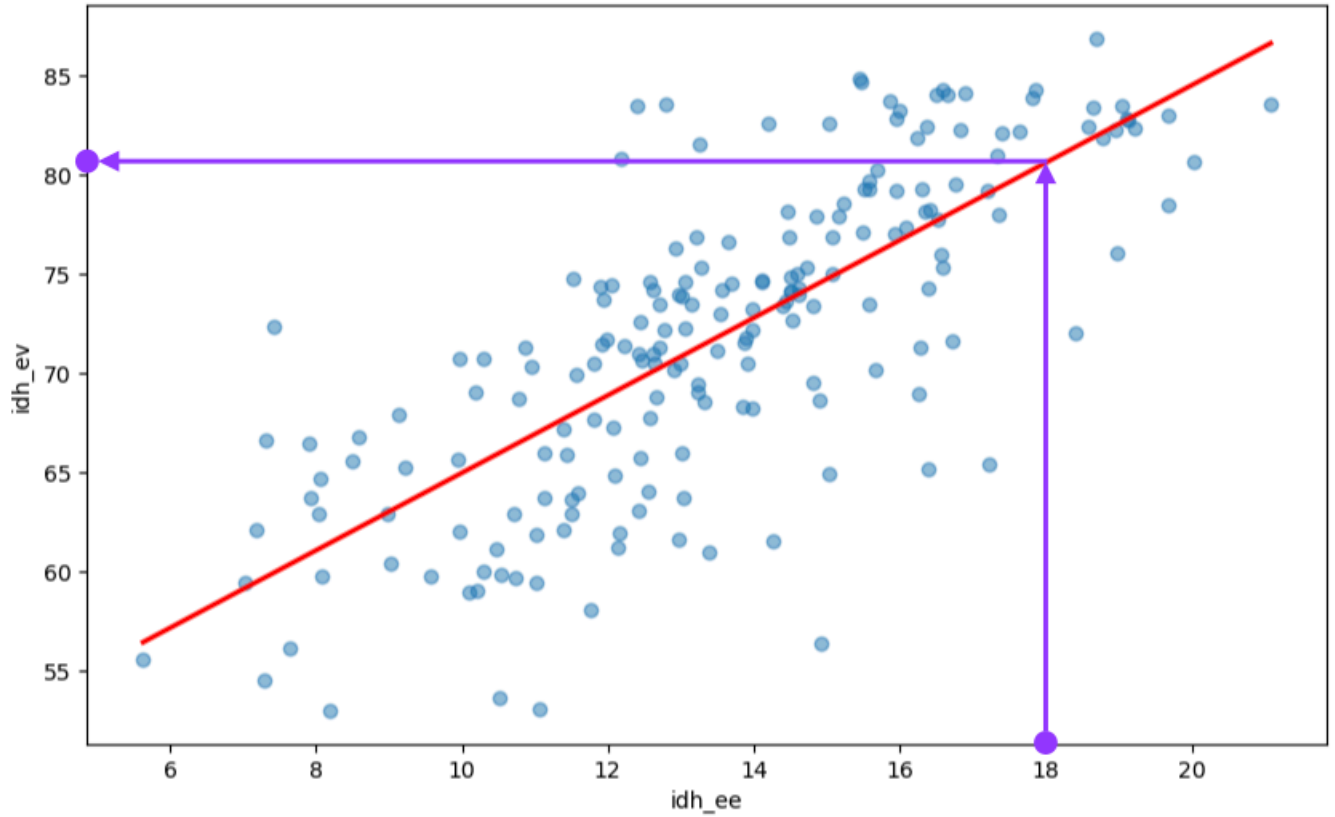
A reta como um modelo



A reta como um modelo



A reta como um modelo



A reta como um modelo

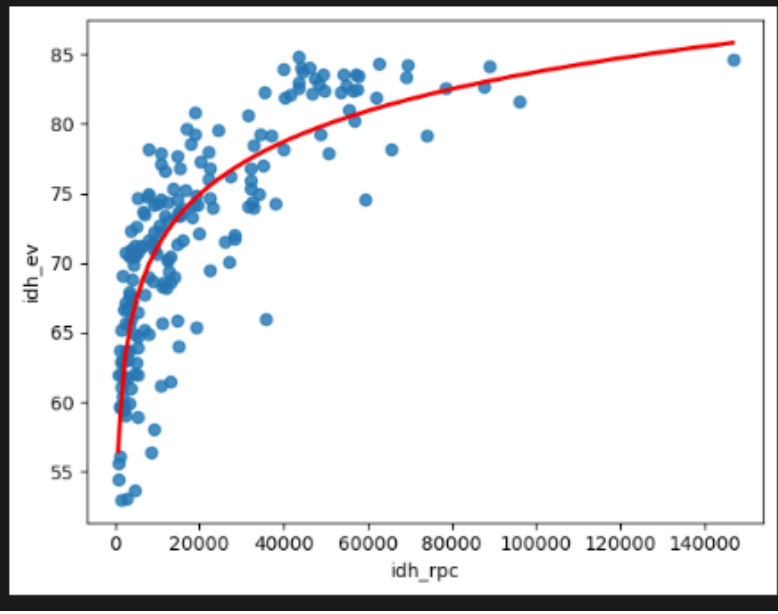
A reta de tendência que encontramos e sua a equação podem ser entendidas também como um **'modelo linear'**, por meio do qual podemos estimar os valores da nossa variável de resposta (expectativa de vida) a partir de variáveis explicativas (nesse caso só temos uma: a expectativa de escolaridade).

Quando você ouvir a expressão "modelo" referindo-se a modelagem matemática, estatística, econômica, etc., não se assuste. **O modelo nada mais é do que uma ou mais equações matemáticas utilizadas para descrever relações entre variáveis e, a partir disso, fazer previsões ou estimativas**.

Scatterplot e regplot: relações entre variáveis numéricas

```
sns.regplot(
    x="idh_rpc",
    y="idh_ev",
    ci=None,
    logx=True,
    line_kws={"color": "red"},
    data=idh_2022
)
```

[17] Python



Primeiro exemplo de anotação

- Vamos destacar a posição do Brasil no nosso scatterplot de Expectativa de vida x Expectativa de escolaridade. Para isso, começamos criando uma nova variável que utilizaremos para colorir o ponto do Brasil.

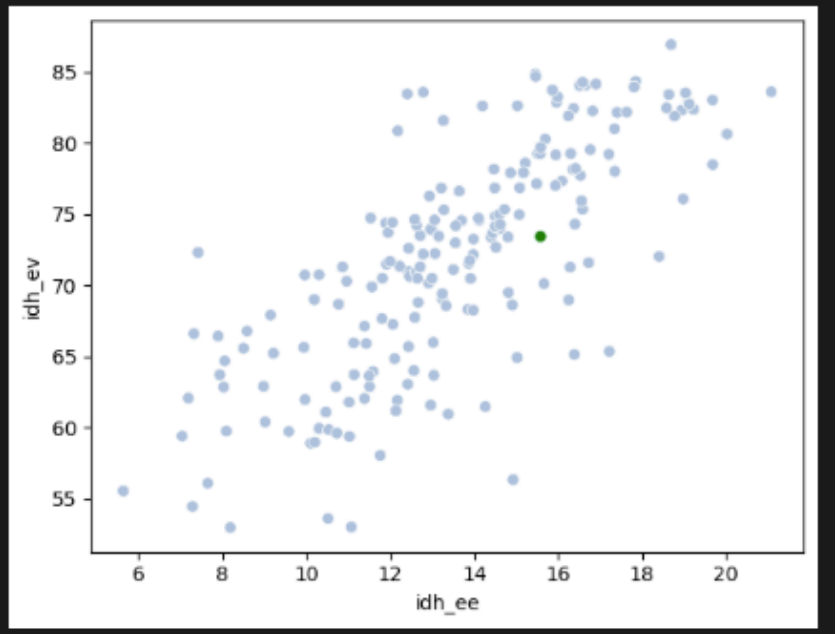
```python
idh_2022["is_br"] = idh_2022.pais == "Brasil"
```

FGV DIREITO RIO

Primeiro exemplo de anotação

```
sns.scatterplot(
    x="idh_ee",
    y="idh_ev",
    hue="is_br",
    palette=["lightsteelblue", "green"],
    legend=False,
    data=idh_2022
)
```

[39] Python



Primeiro exemplo de anotação

Precisamos agora das coordenadas do Brasil no gráfico, para adicionar uma anotação próxima ao ponto. O código abaixo obtém essas coordenadas, acrescidas de um pequeno espaço para evitar sobreposições.

```
[44]
idh_2022_br = idh_2022.query("pais == 'Brasil'")

x_position = idh_2022_br["idh_ee"].values[0] + 0.2
y_position = idh_2022_br["idh_ev"].values[0] + 0.5
```

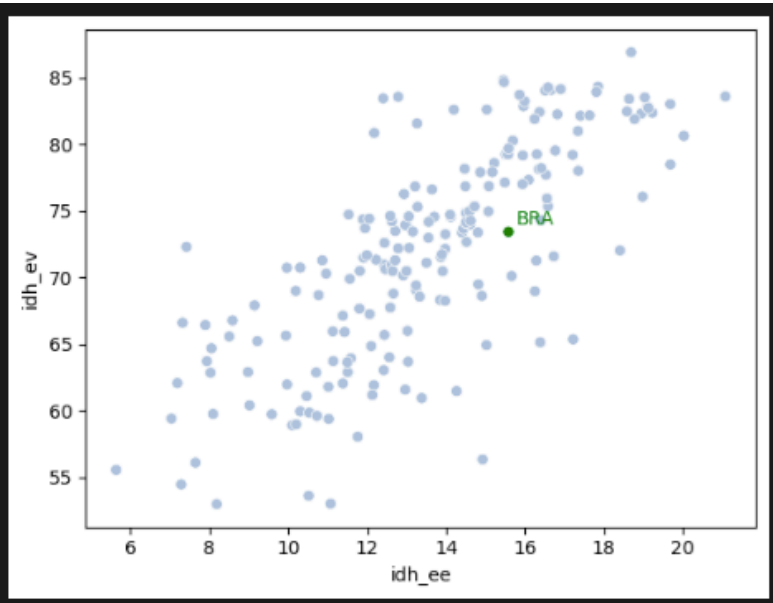
Python

FGV DIREITO RIO

Primeiro exemplo de anotação

```
sns.scatterplot(
    x="idh_ee",
    y="idh_ev",
    hue="is_br",
    palette=["lightsteelblue", "green"],
    legend=False,
    data=idh_2022
)
plt.text(x=x_position, y=y_position, s="BRA", color="green")
```

[46] Python



Scatterplot e regplot: relações entre variáveis numéricas

```python
fig, ax = plt.subplots(figsize=(10, 6))

sns.scatterplot(
    x="idh_ee",
    y="idh_ev",
    hue="is_br",
    palette=["lightsteelblue", "green"],
    legend=False,
    data=idh_2022,
    ax=ax
)

plt.text(x=x_position, y=y_position, s="BRA", color="green")

ax.set_title("Expectativa de vida x Expectativa de escolaridade (2022)")
ax.set_xlabel("Expectativa de escolaridade")
ax.set_ylabel("Expectativa de vida")

plt.show()
```

[52] Python

Mãos à obra!