FGV DIREITO RIO

Explorando distribuições – KDE, ECDF e Boxplot

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

- Gráficos KDE e ECDF
- Quartis e intervalo interquartil
- Boxplot
- Query

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

KDE (Kernel Density Estimate)

- Como podemos criar um gráfico para compreender a distribuição dos valores do IDH dos países em 2022 (coluna `"idh"` da nossa base) usando a função `kdeplot`?

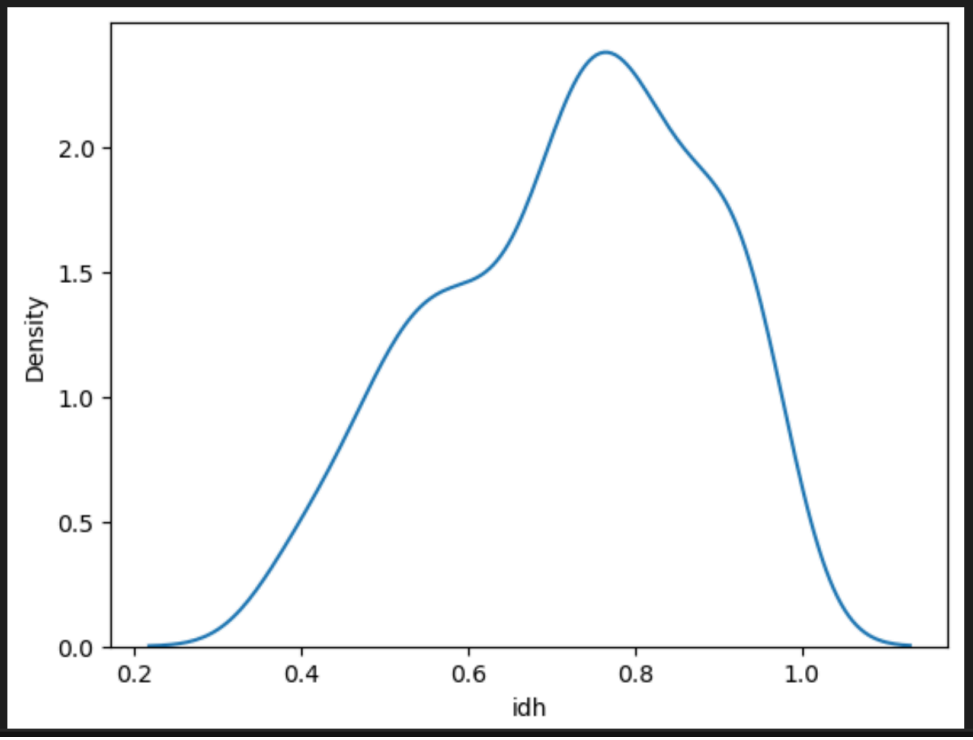
KDE (Kernel Density Estimate)

- Como podemos criar um gráfico para compreender a distribuição dos valores do IDH dos países em 2022 (coluna `"idh"` da nossa base) usando a função `kdeplot`?
    - Você consegue ver algum problema nesse gráfico?

```python
sns.kdeplot(x="idh", data=idh_2022)
```

Python

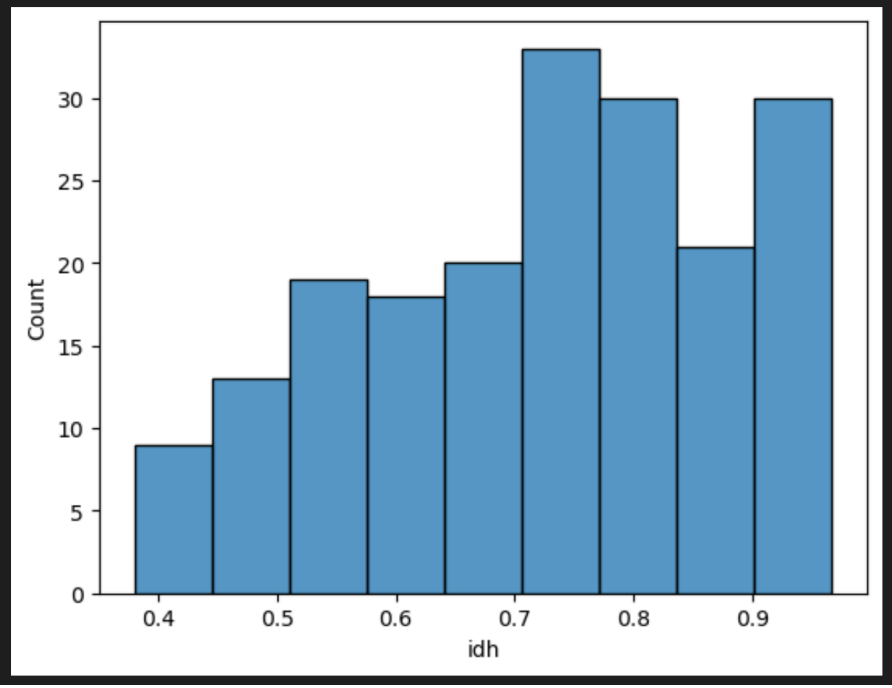
```
<Axes: xlabel='idh', ylabel='Density'>
```



```python
sns.histplot(x="idh", data=idh_2022)
```

Python

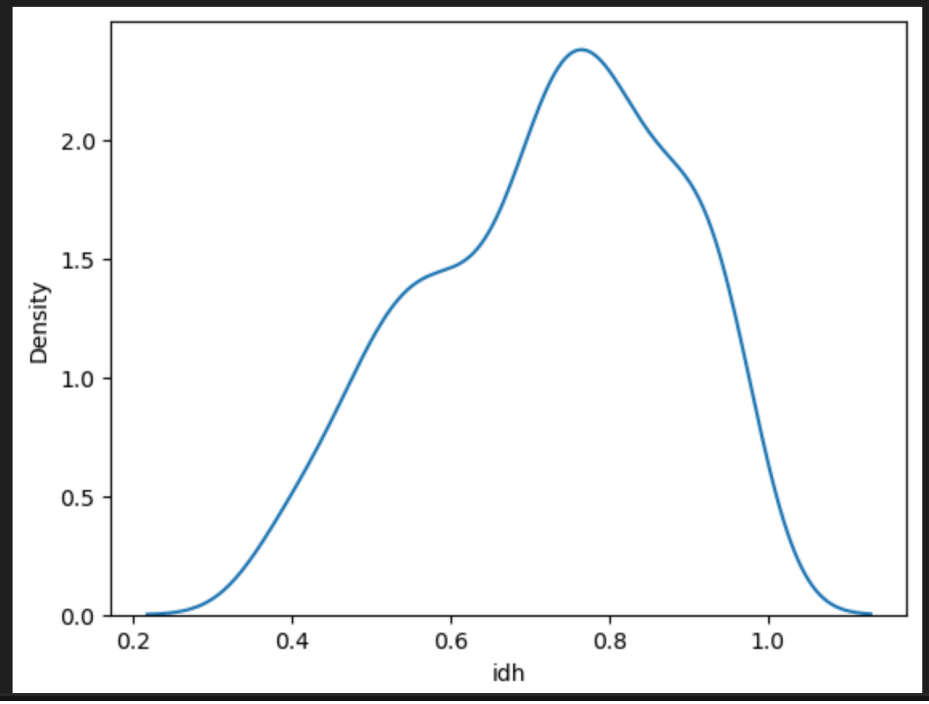
```
<Axes: xlabel='idh', ylabel='Count'>
```




```python
sns.kdeplot(x="idh", data=idh_2022)
```

Python

```
<Axes: xlabel='idh', ylabel='Density'>
```



ECDF (Empirical Cum. Dist. Function)

- Como podemos criar um gráfico para compreender a distribuição dos valores do IDH dos países em 2022 (coluna `"idh"` da nossa base) usando a função `ecdfplot`?

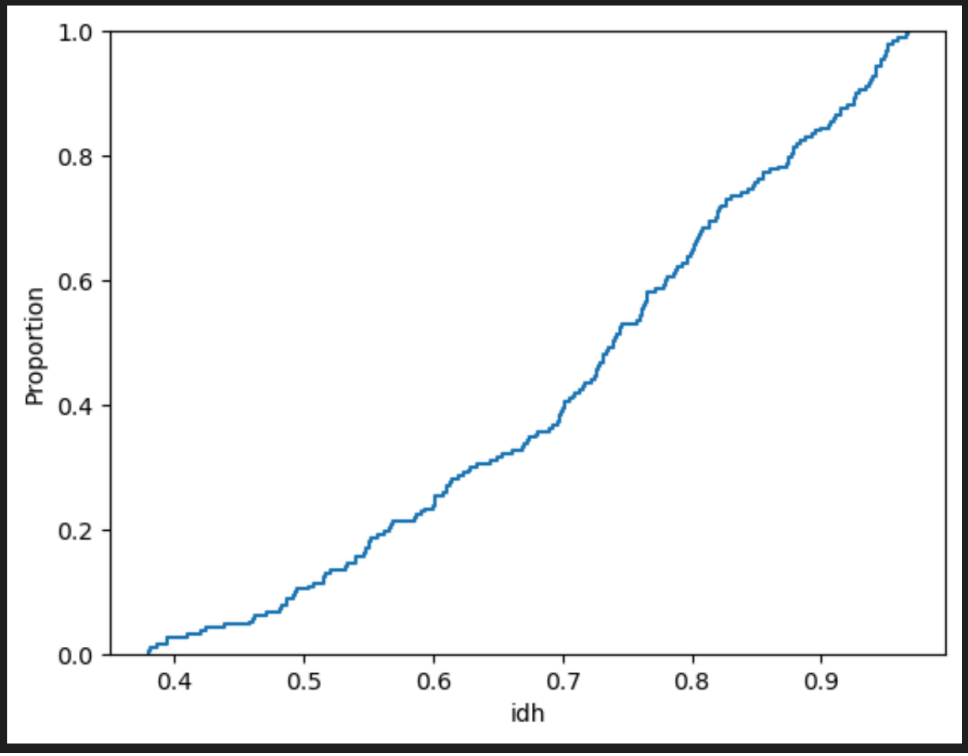
ECDF (Empirical Cum. Dist. Function)

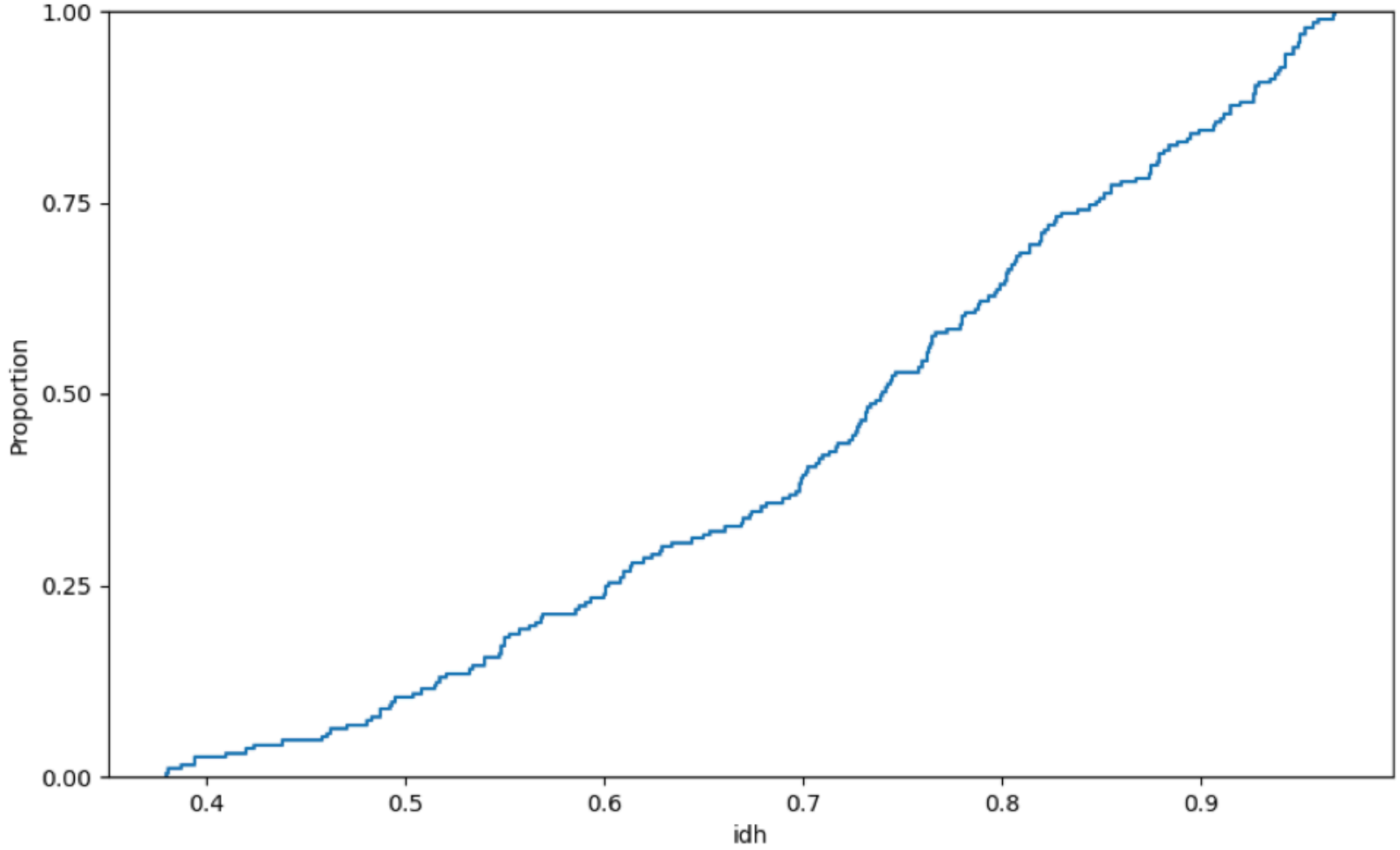
- Como podemos criar um gráfico para compreender a distribuição dos valores do IDH dos países em 2022 (coluna `"idh"` da nossa base) usando a função `ecdfplot`?
    - Como interpretamos esse gráfico e quais são as suas limitações?

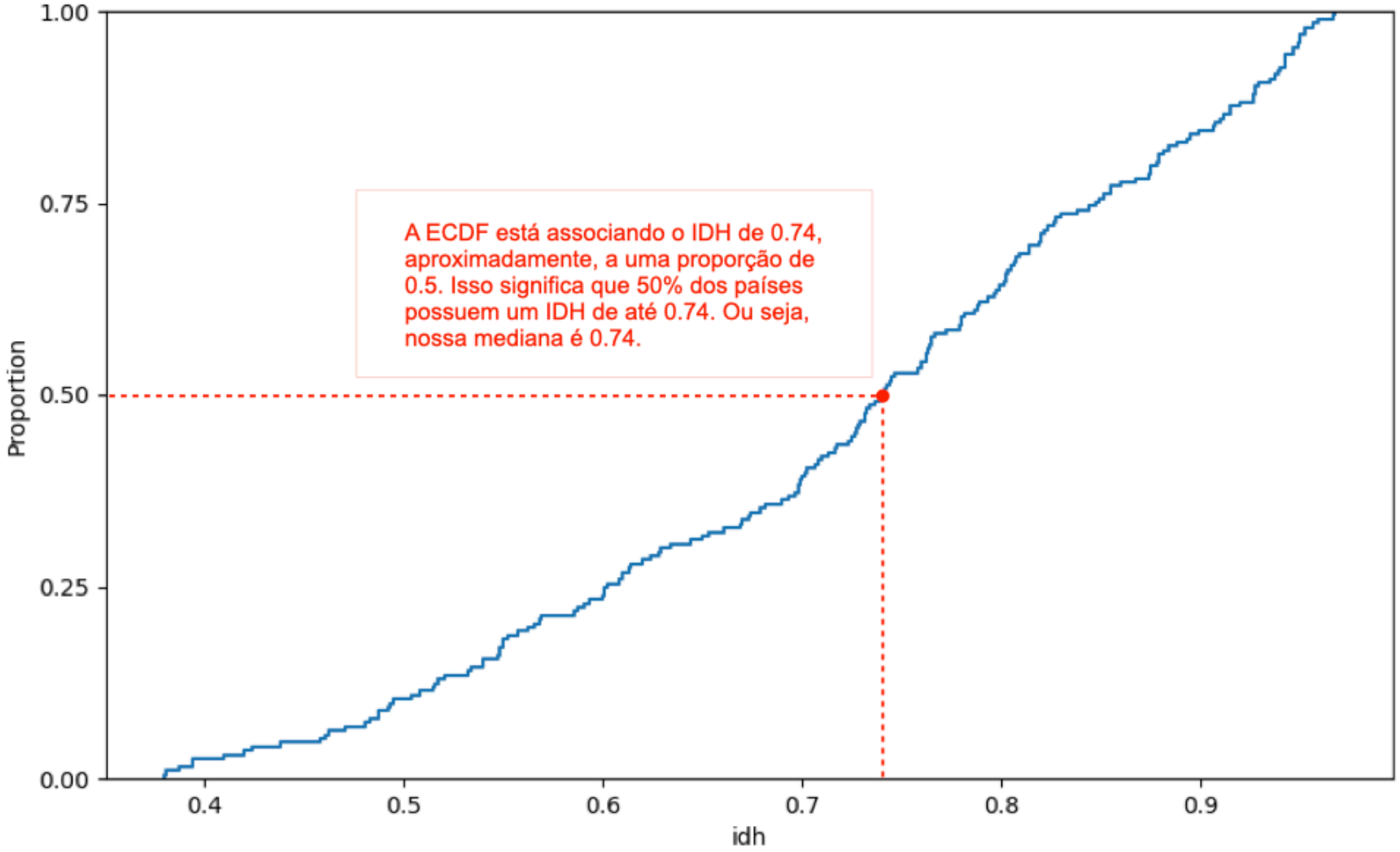
```python
sns.ecdfplot(x="idh", data=idh_2022)
```

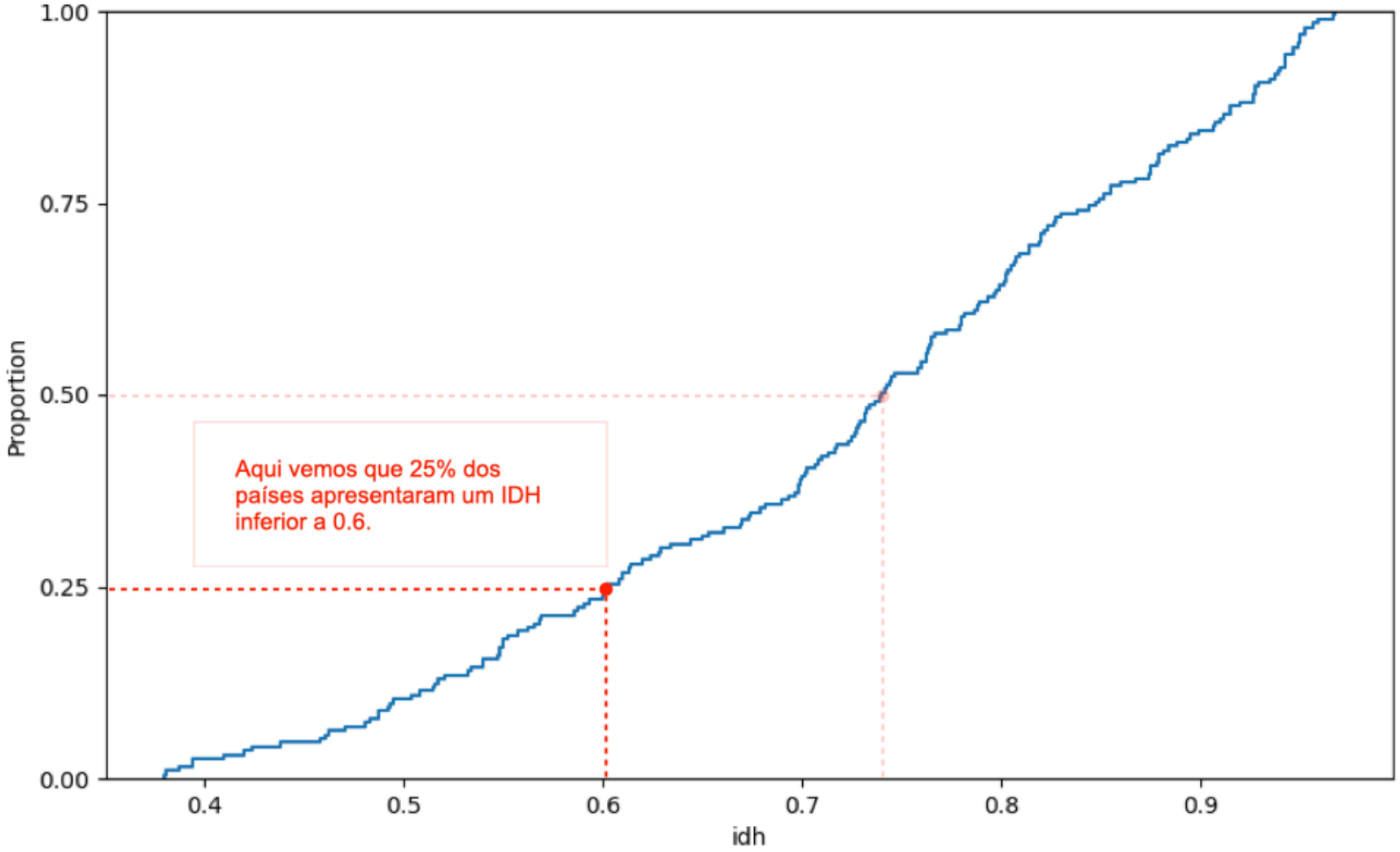
Python

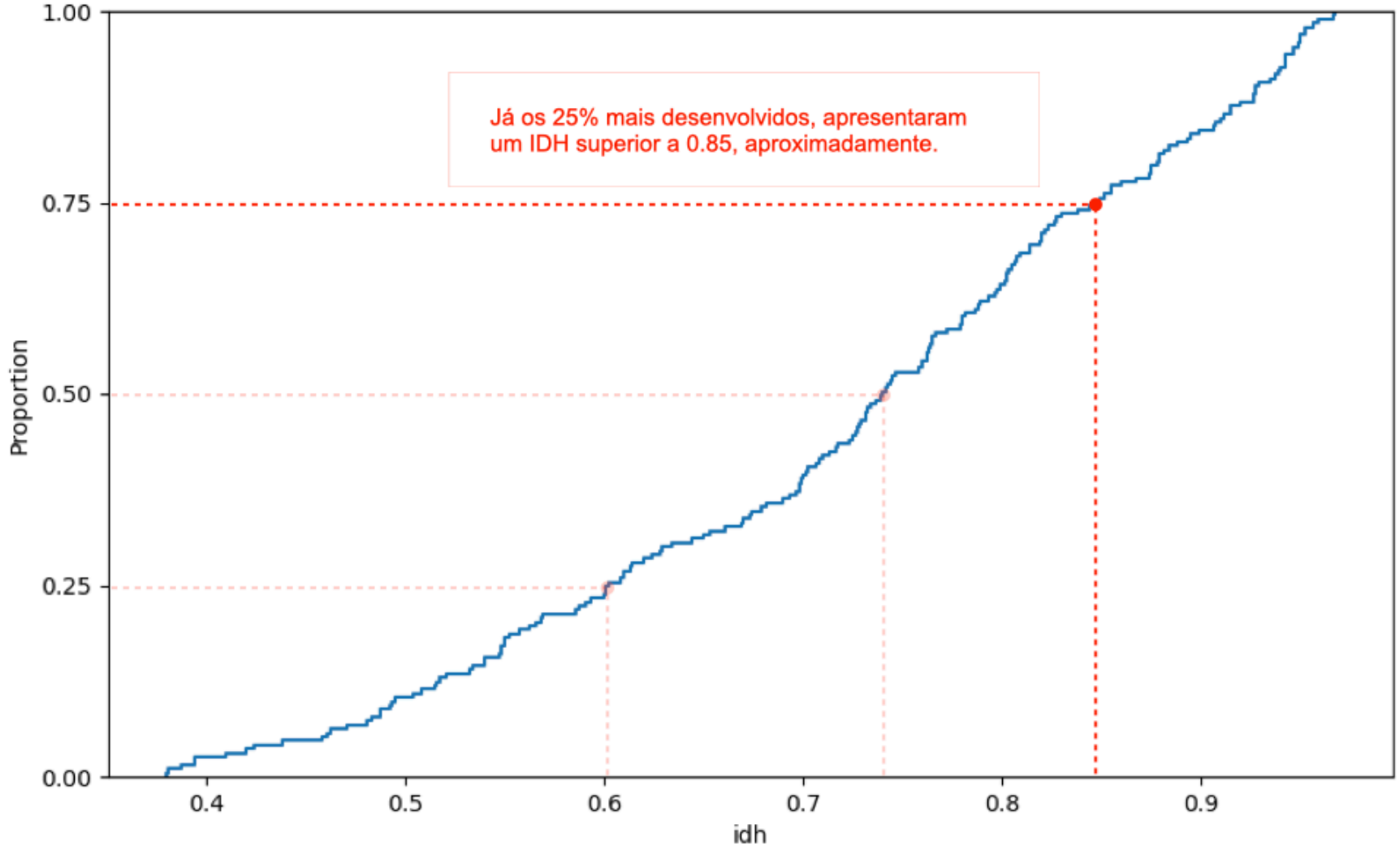
```
<Axes: xlabel='idh', ylabel='Proportion'>
```

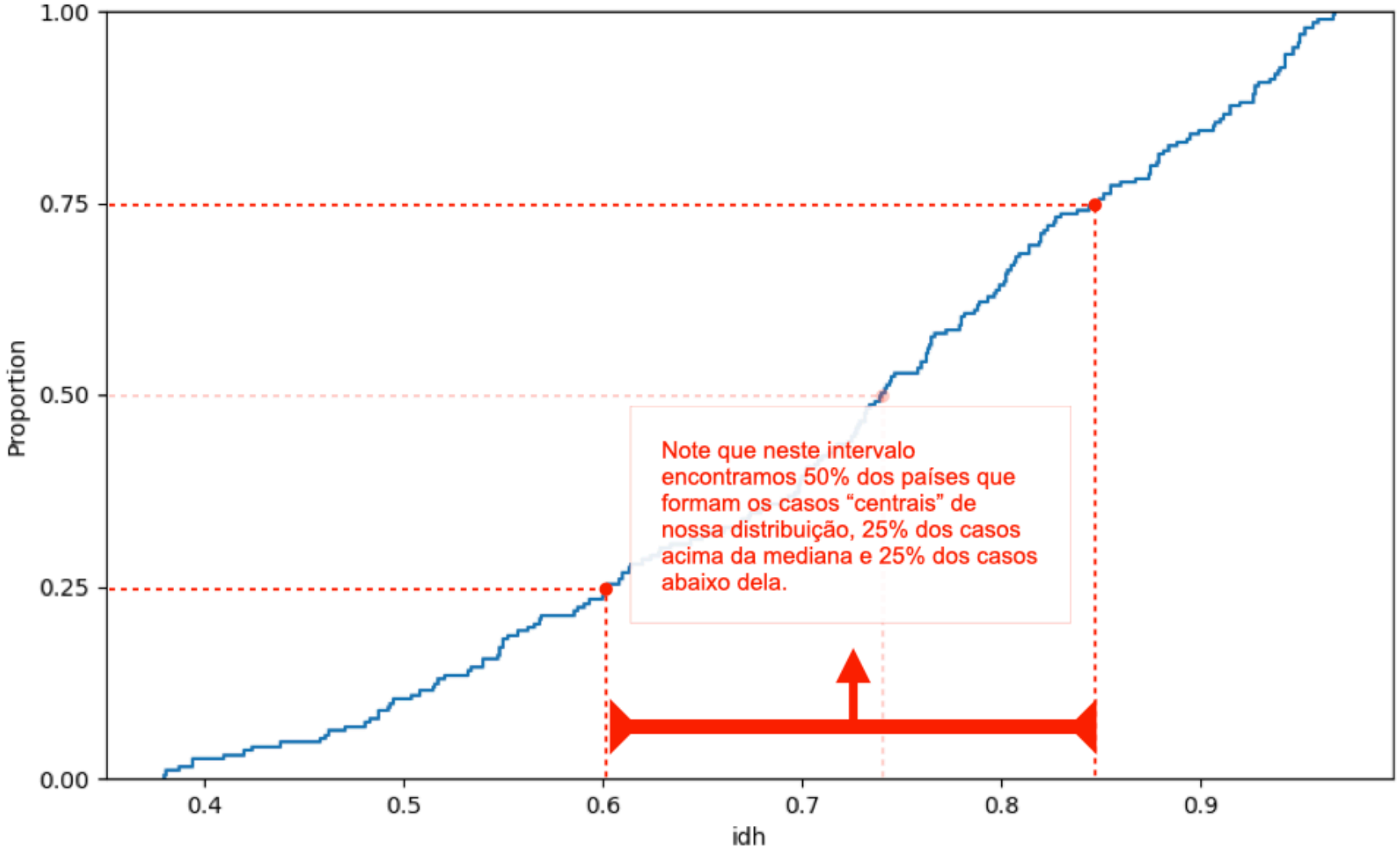


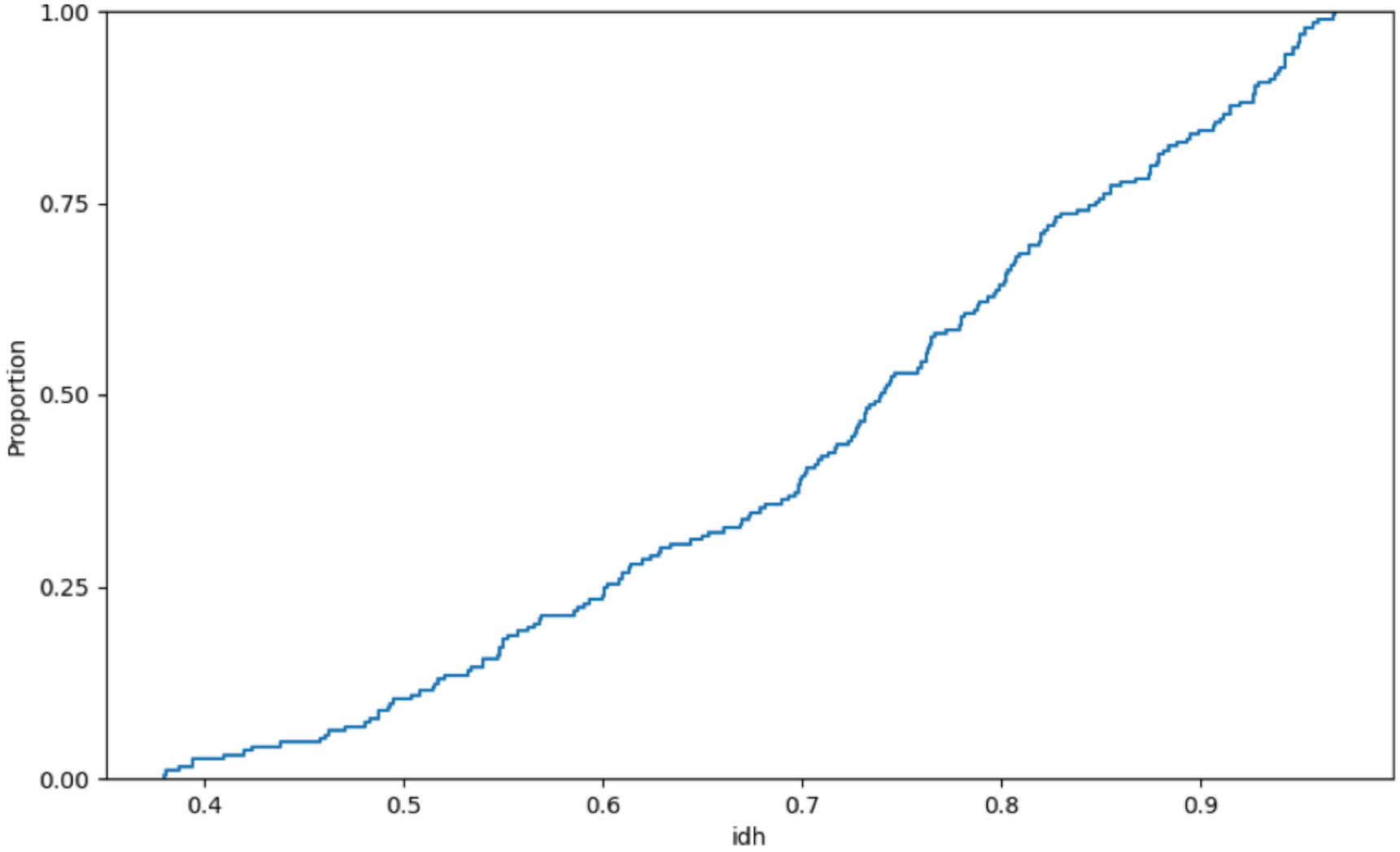

1.00
0.75
0.50
0.25
0.00
Proportion
0.4
0.5
0.6
0.7
0.8
0.9
idh


1.00
0.75
0.50
0.25
0.00
Proportion
A ECDF está associando o IDH de 0.74, aproximadamente, a uma proporção de 0.5. Isso significa que 50% dos países possuem um IDH de até 0.74. Ou seja, nossa mediana é 0.74.
0.4
0.5
0.6
0.7
0.8
0.9
idh


1.00
0.75
0.50
0.25
0.00
Proportion
Aqui vemos que 25% dos países apresentaram um IDH inferior a 0.6.
0.4
0.5
0.6
0.7
0.8
0.9
idh


Já os 25% mais desenvolvidos, apresentaram
um IDH superior a 0.85, aproximadamente.
1.00
0.75
0.50
0.25
0.00
Proportion
0.4
0.5
0.6
0.7
0.8
0.9
idh


1.00
0.75
0.50
0.25
0.00
Proportion
0.4
0.5
0.6
0.7
0.8
0.9
idh
Note que neste intervalo encontramos 50% dos países que formam os casos “centrais” de nossa distribuição, 25% dos casos acima da mediana e 25% dos casos abaixo dela.


1.00
0.75
0.50
0.25
0.00
Proportion
0.4
0.5
0.6
0.7
0.8
0.9
idh

A divisão de uma distribuição em quartis

- Vamos trabalhar com as idades de 12 pessoas.
- `Idade`: 22, 24, 25, 28, 30, 35, 40, 42, 45, 50, 54, 60.
- **Quartis como "partes" da distribuição de valores**:
  - Agora vamos dividir essa distribuição em quatro partes:

- **1º Quartil**
  - 22
  - 24
  - 25
- **2º Quartil**
  - 28
  - 30
  - 35
- **3º Quartil**
  - 40
  - 42
  - 45
- **4º Quartil**
  - 50
  - 54
  - 60

A divisão de uma distribuição em quartis

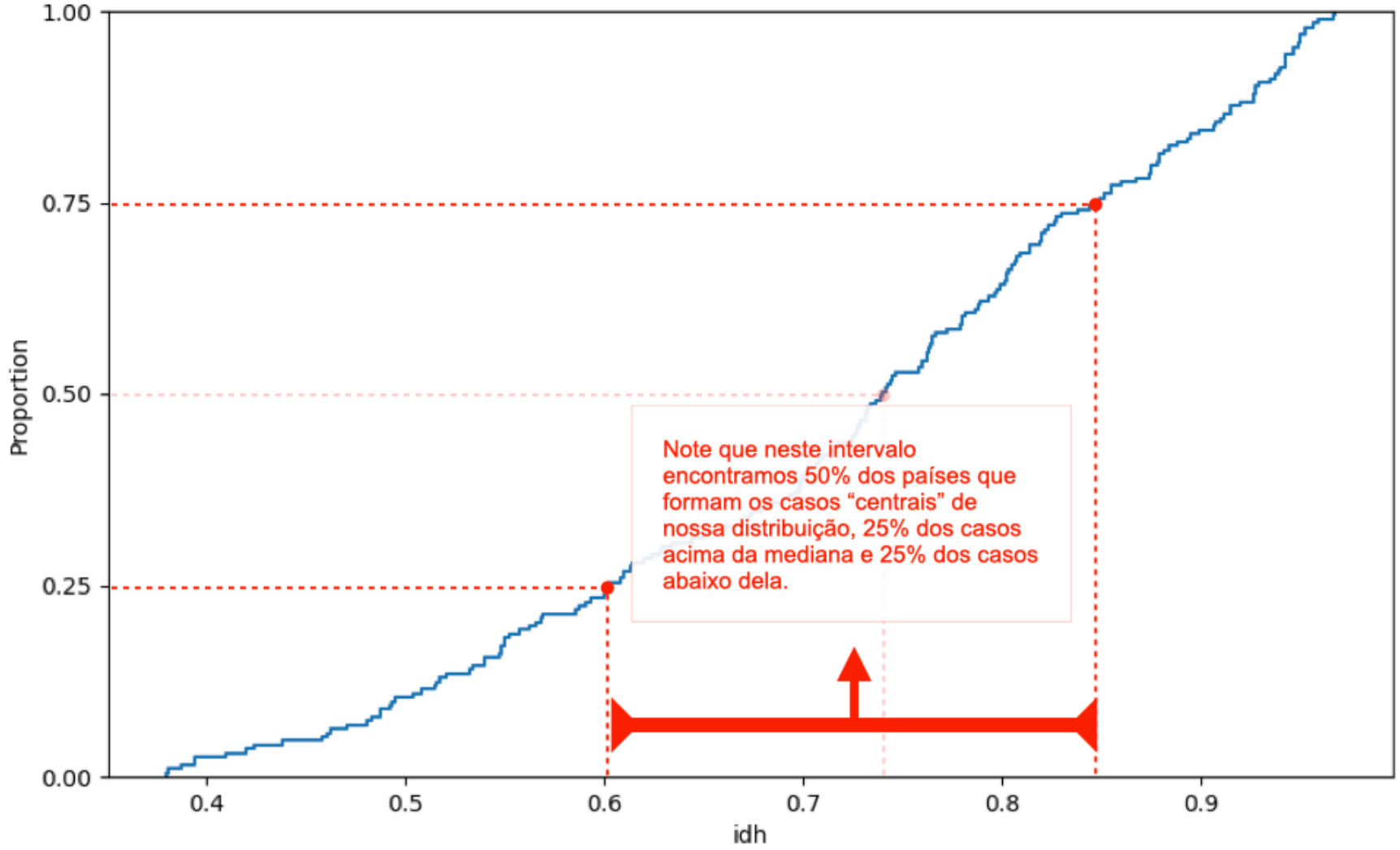
- `Idade`: 22, 24, 25, 28, 30, 35, 40, 42, 45, 50, 54, 60.
- **Quartis como fronteiras entre as partes**:
  - Por vezes usamos o termo "quartil" para designar as **3 fronteiras** entre as partes da nossa distribuição:
    - **Q1**: divide o 1º quartil do 2º quartil. Ou seja, 25% dos casos estão abaixo desse valor.
    - **Q2**: divide o 2º quartil do 3º quartil. Abaixo desse valor teríamos metade dos casos, ou seja, **Q2 é a mediana**.
    - **Q3**: divide o 2º quartil do 3º quartil. Ou seja, 75% dos casos estão abaixo desse valor (e 25% estão acima).

Intervalo Interquartil

- O intervalo interquartil (IQR) é dado pela distância entre $Q1$ e $Q3$. Ou seja:

$$IQR = Q3 - Q1$$

- O IQR é importante porque ele nos dá uma medida do grau de dispersão da nossa distribuição, tomando como base os casos "centrais", ou seja, os valores que estão em torno da nossa mediana.

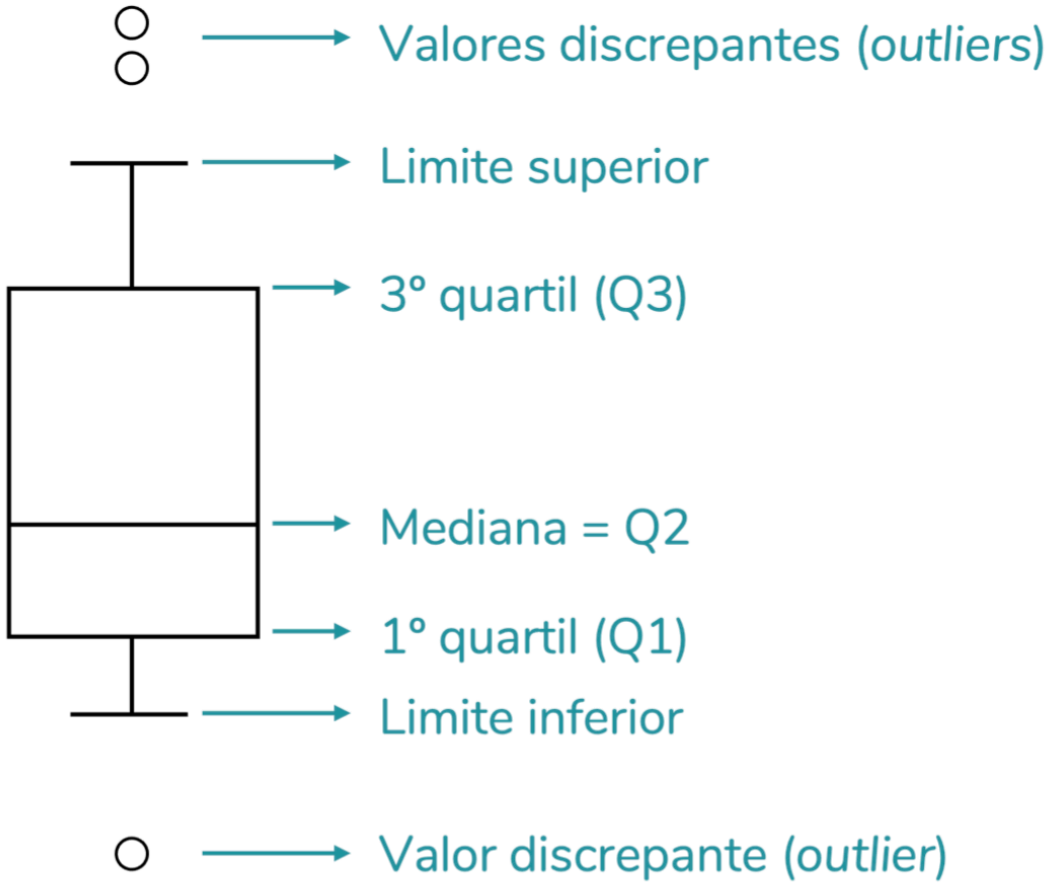

1.00
0.75
0.50
0.25
0.00
Proportion
0.4
0.5
0.6
0.7
0.8
0.9
idh
Note que neste intervalo encontramos 50% dos países que formam os casos “centrais” de nossa distribuição, 25% dos casos acima da mediana e 25% dos casos abaixo dela.

Boxplot

- O gráfico boxplot oferece muitas informações sobre a distribuição, de forma condensada.
- Ele é um gráfico extremamente informativo, desde que se saiba lê-lo.

Boxplot

- O gráfico boxplot oferece muitas informações sobre a distribuição, de forma condensada.
- Ele é um gráfico extremamente informativo, desde que se saiba lê-lo.
- A figura ao lado indica os significados associados a cada dimensão da caixa de um Boxplot.



Boxplot

- Como podemos criar um gráfico para compreender a distribuição dos valores do IDH dos países em 2022 (coluna `"idh"` da nossa base) usando a função `boxplot`?

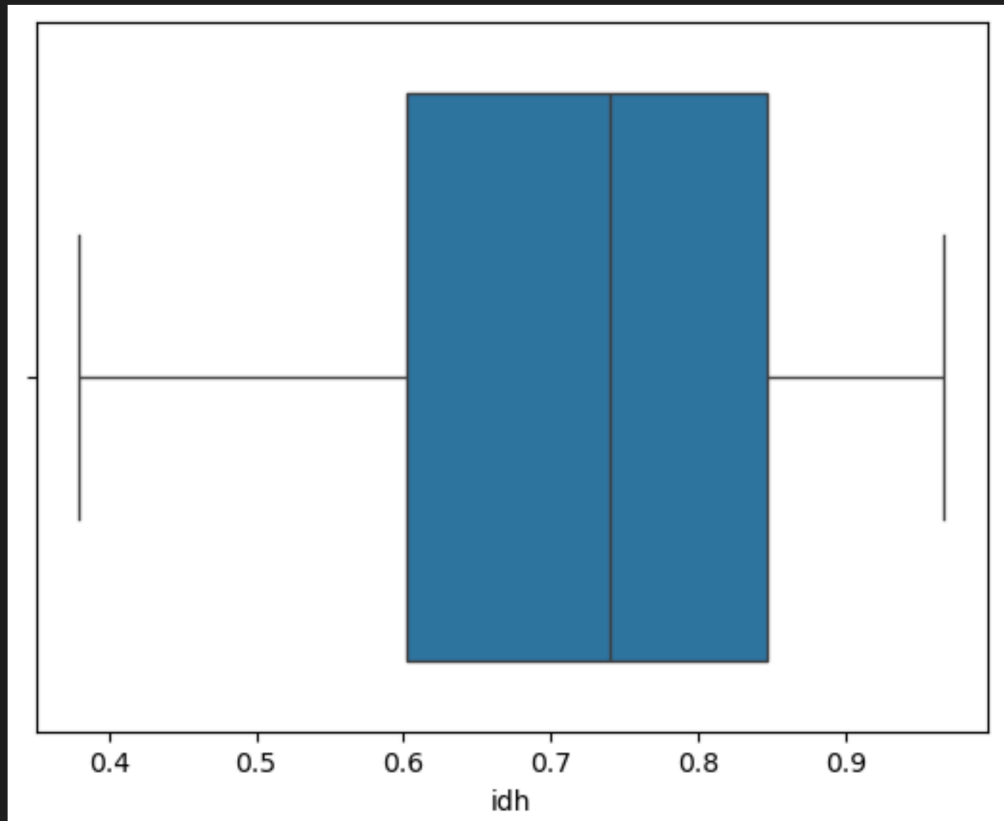
Boxplot

- Como podemos criar um gráfico para compreender a distribuição dos valores do IDH dos países em 2022 (coluna `"idh"` da nossa base) usando a função `boxplot`?

```
sns.boxplot(x="idh", data=idh_2022)
```

✓ 0.0s Python

```
<Axes: xlabel='idh'>
```

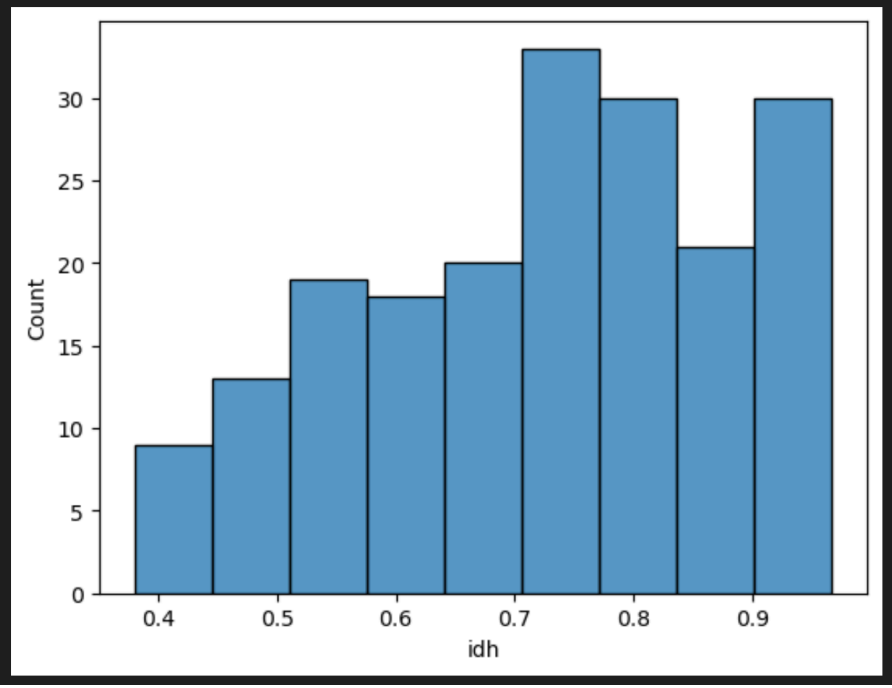


Explorando distribuições – KDE, ECDF e Boxplot

```
sns.histplot(x="idh", data=idh_2022)
```

Python

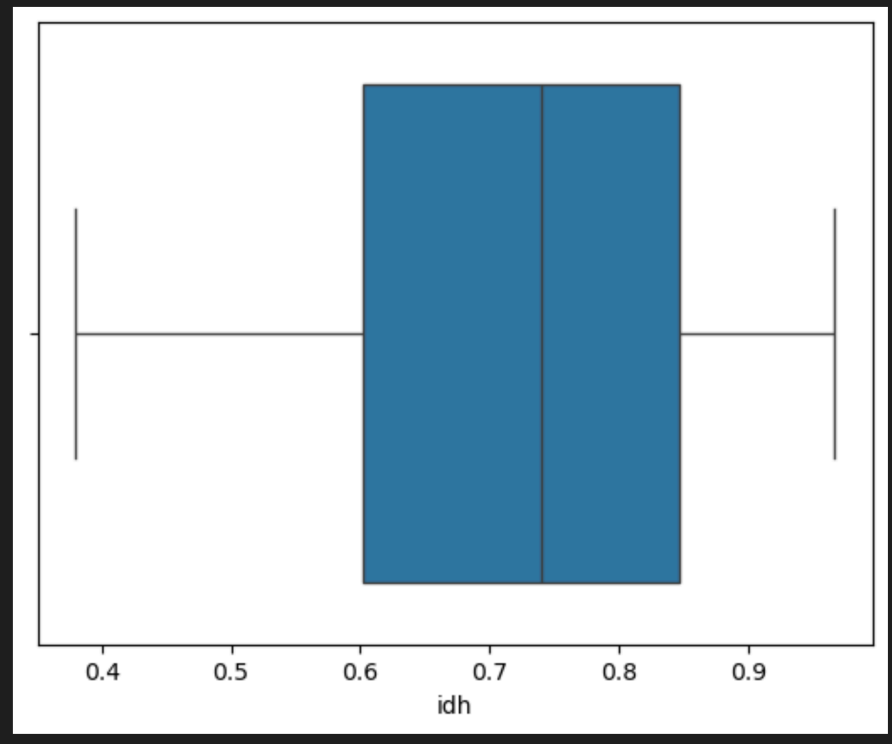
```
<Axes: xlabel='idh', ylabel='Count'>
```



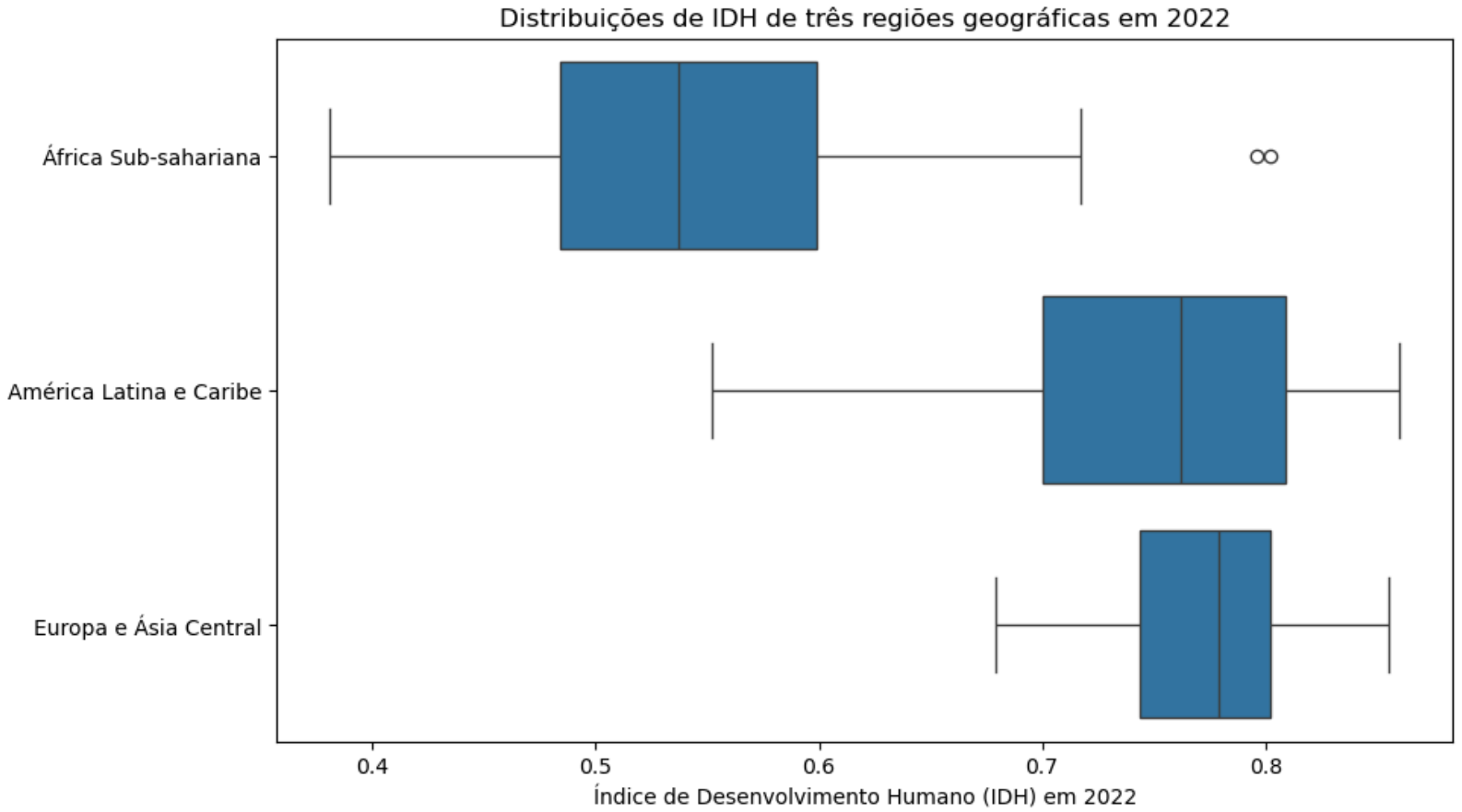

```
sns.boxplot(x="idh", data=idh_2022)
```

✓ 0.0s Python

```
<Axes: xlabel='idh'>
```



O boxplot é muito útil para comparar distribuições!



Filtrando dados com o método `.query()`

```python
regioes_selecionadas = [
    "África Sub-sahariana",
    "América Latina e Caribe",
    "Europa e Ásia Central"
]

idh_regioes_selecionadas = idh_2022.query("regiao in @regioes_selecionadas")

sns.boxplot(
    x="idh",
    y="regiao",
    order=regioes_selecionadas,
    data=idh_regioes_selecionadas
)
```

Python

Mãos à obra!