Python para Análise de Dados

Quem sou ...

- **Andrius Jaques**
- **Alegretense**
- **Programador e Analista de Dados**
- **Entusiasta Dados Abertos**

Sobre o que vamos conversar

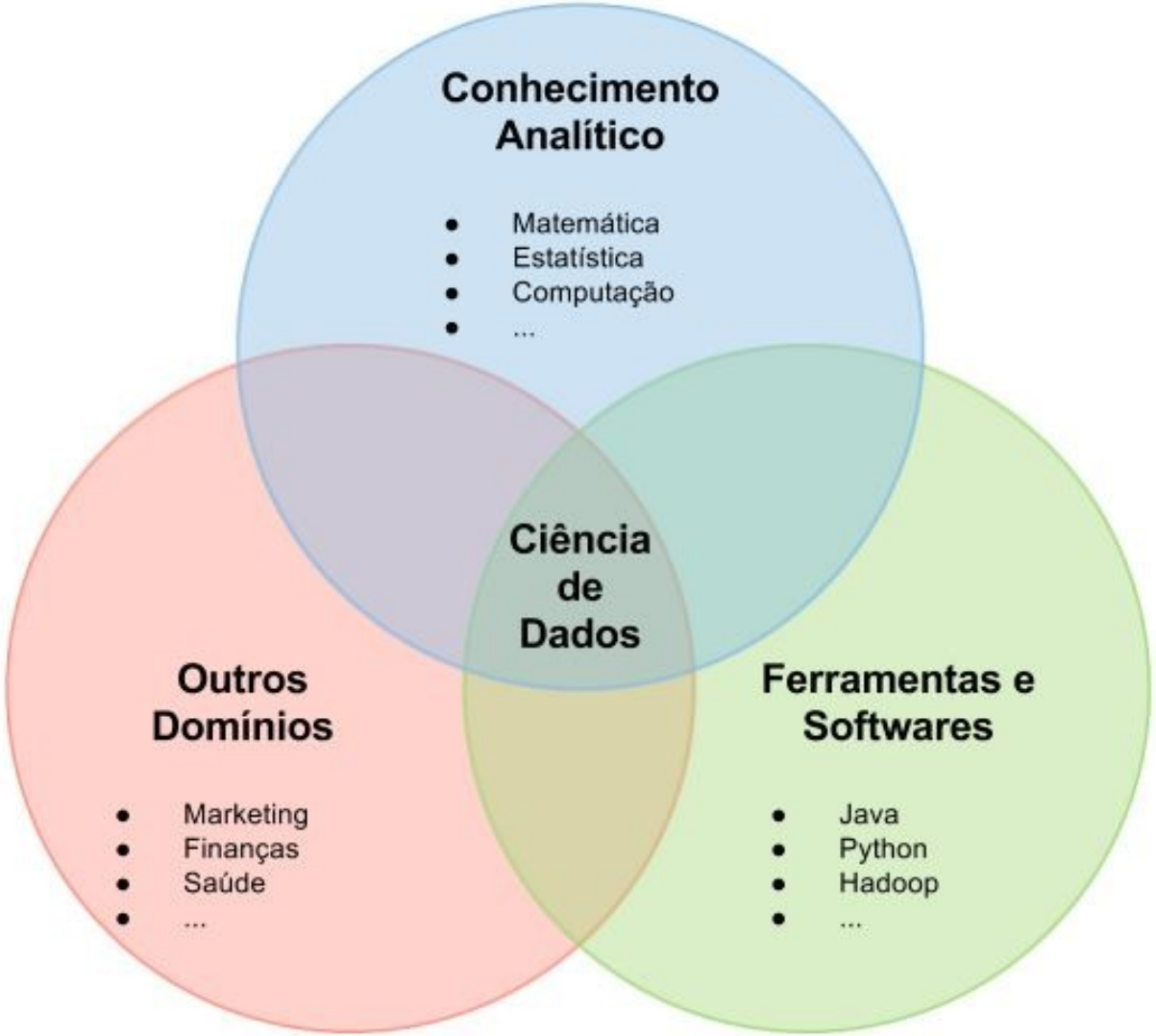
- Análise de dados, ciência de dados
- Bibliotecas análise de dados:
- Explorando PANDAS
- Perguntas

Análise e Ciência de Dados

Análise de Dados

- **Transformar dados em informação**
- **Tratar e limpar dados**
- **Explorar dados:**
  - padrões
  - tedências;

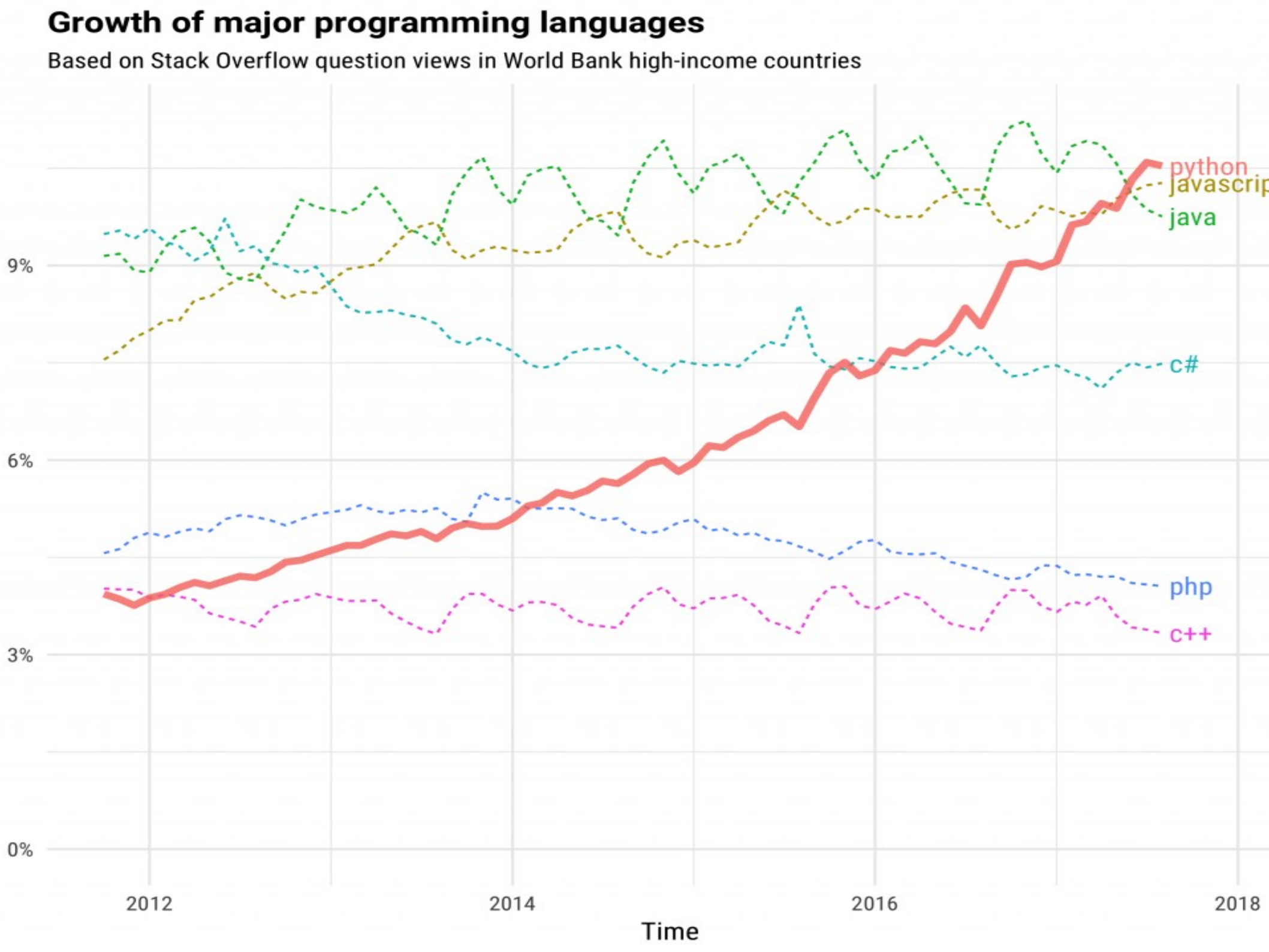
Ciência de Dados



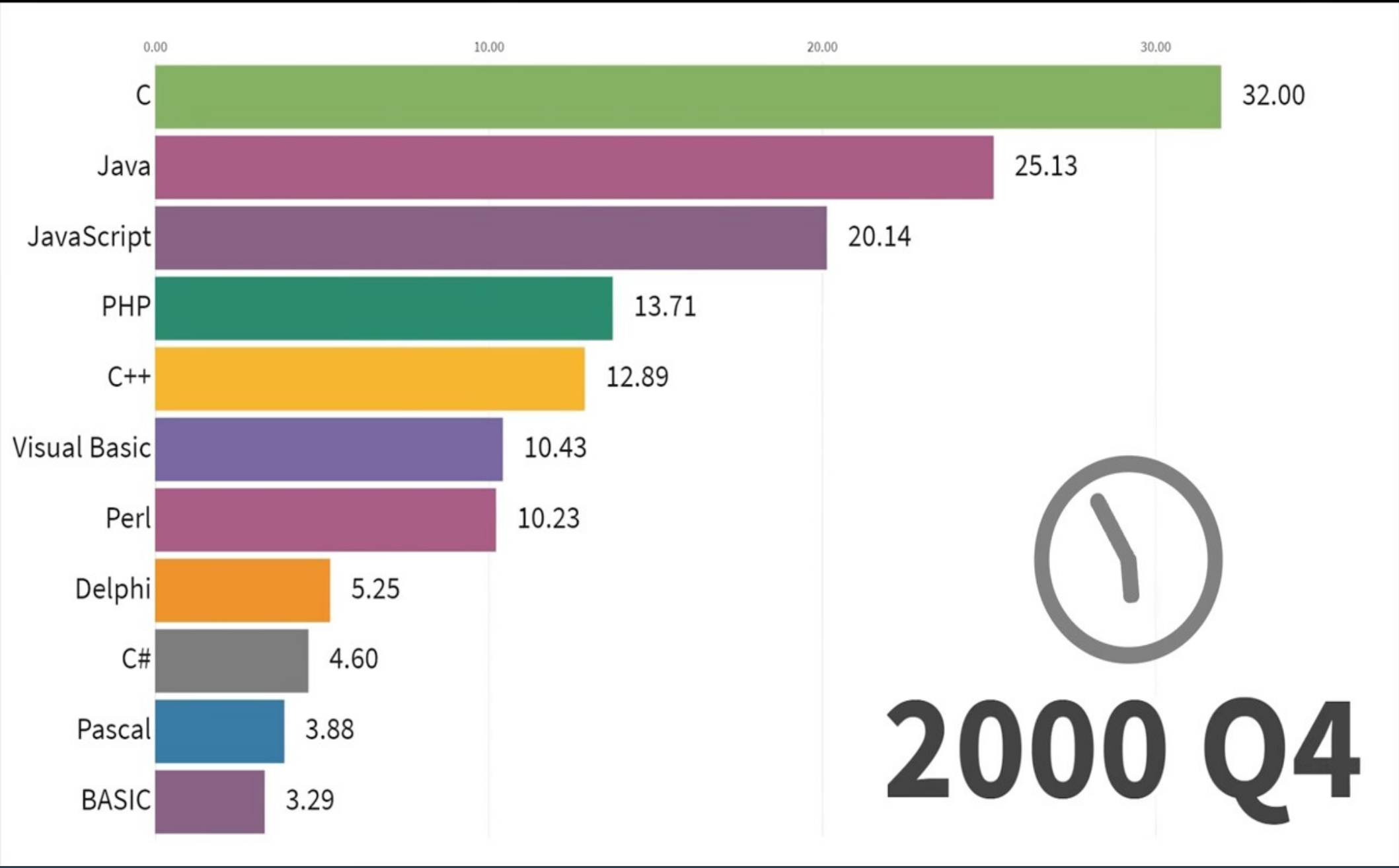
Ciência de Dados

Requisitos: python para análise de dados

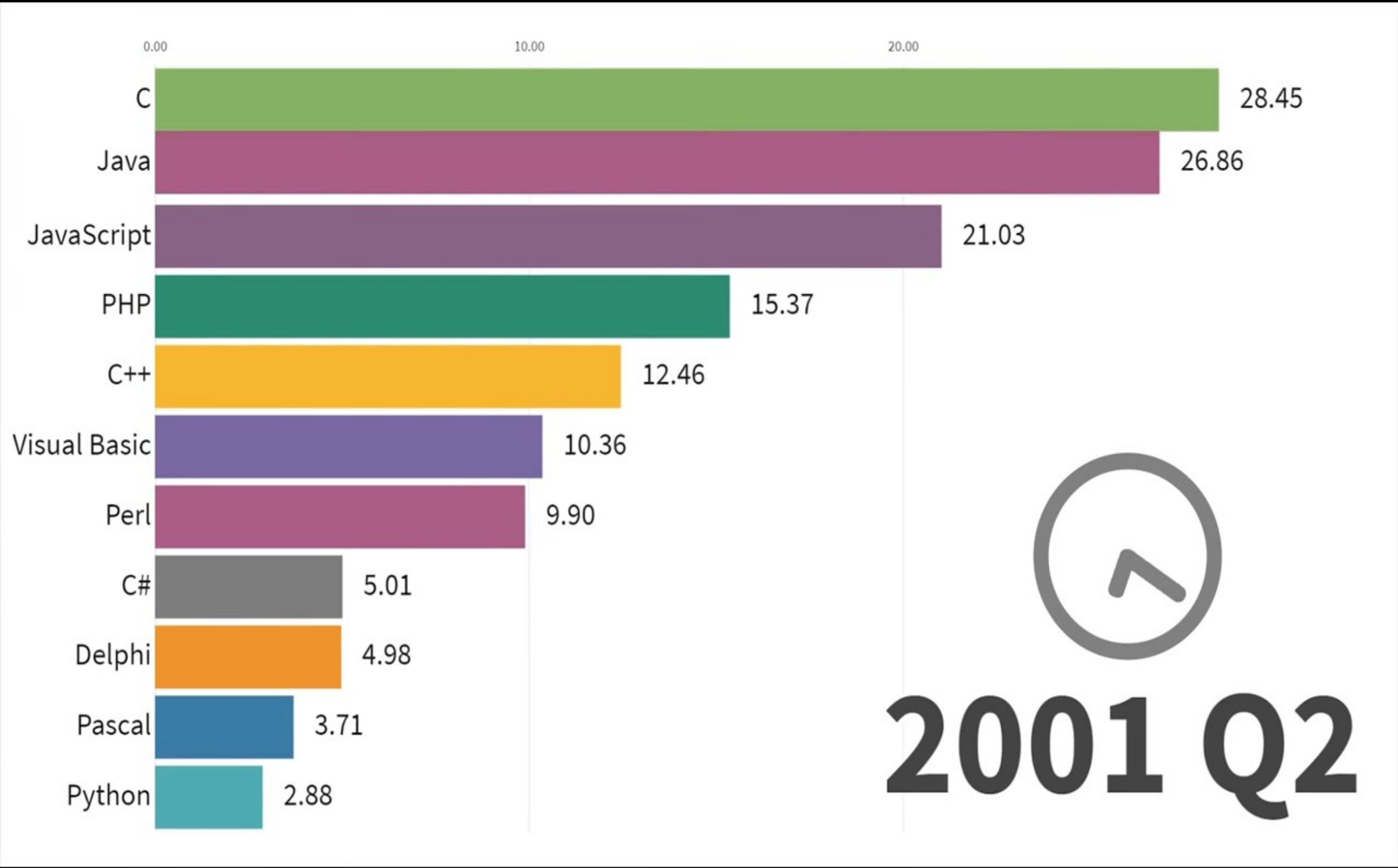
Crescimento da linguagem python


Growth of major programming languages
Based on Stack Overflow question views in World Bank high-income countries
python
javascrip
java
c#
php
c++
9%
6%
3%
0%
2012
2014
2016
2018
Time

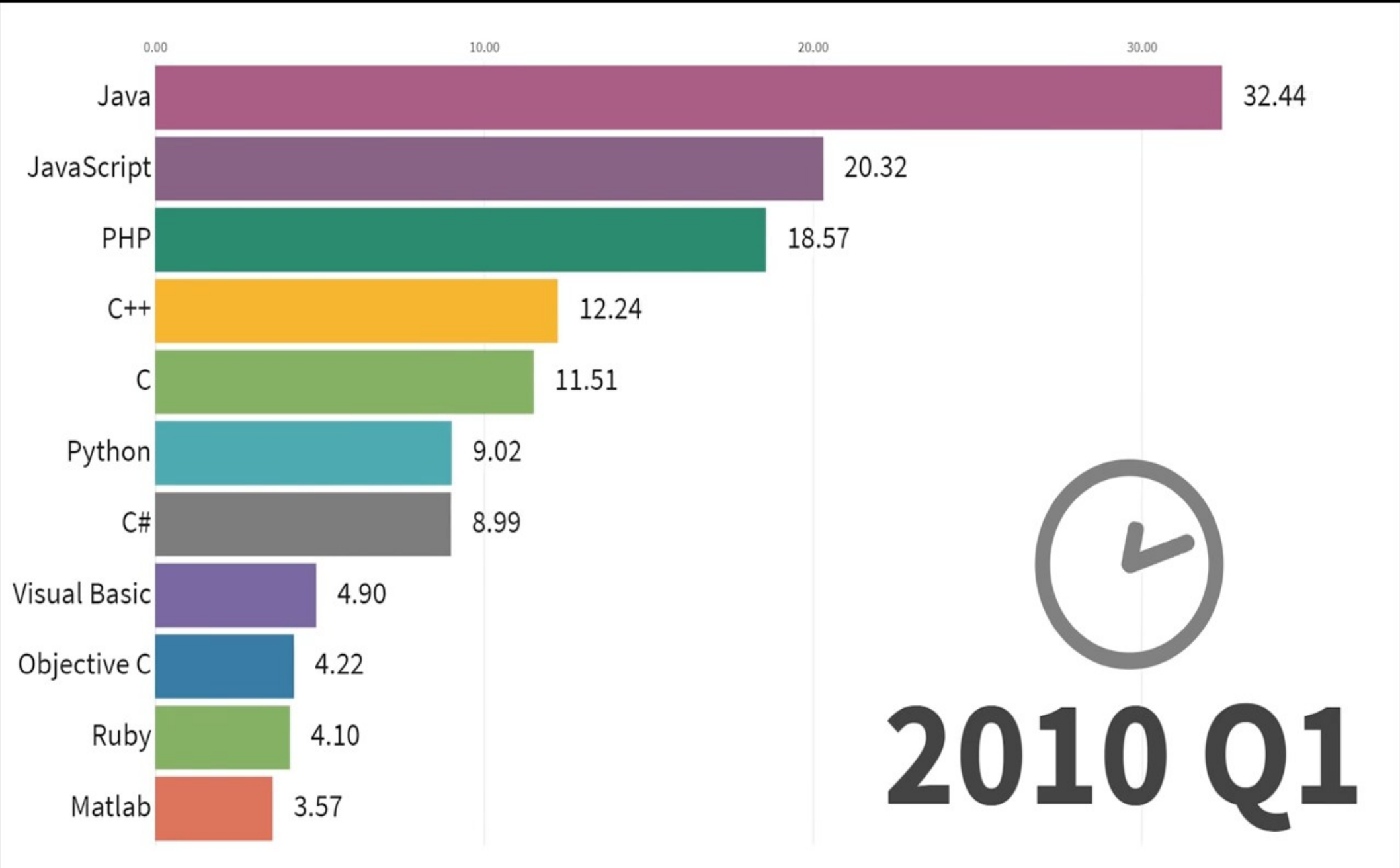
Most Popular Programming Languages 1965 - 2019



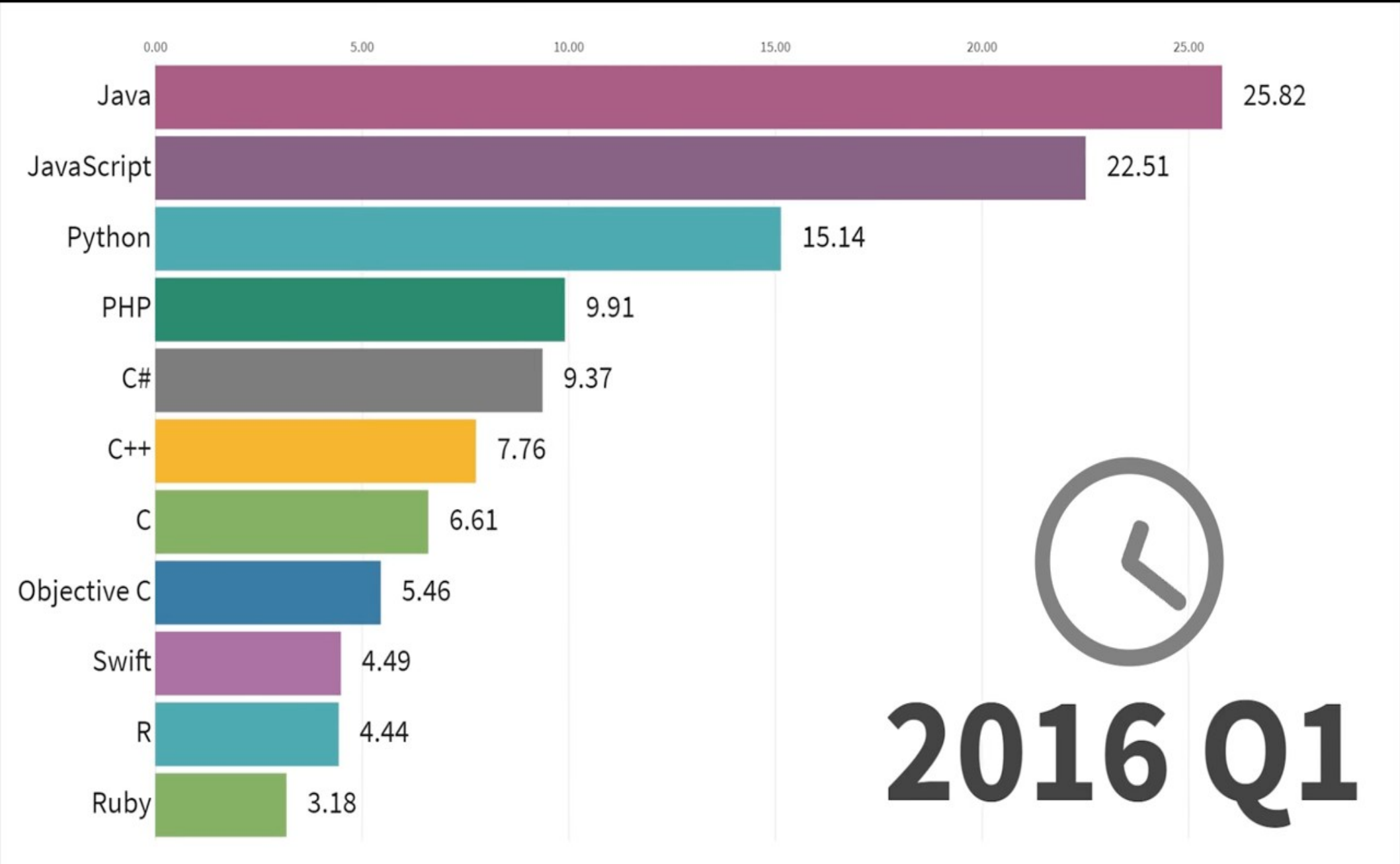
Most Popular Programming Languages 1965 - 2019



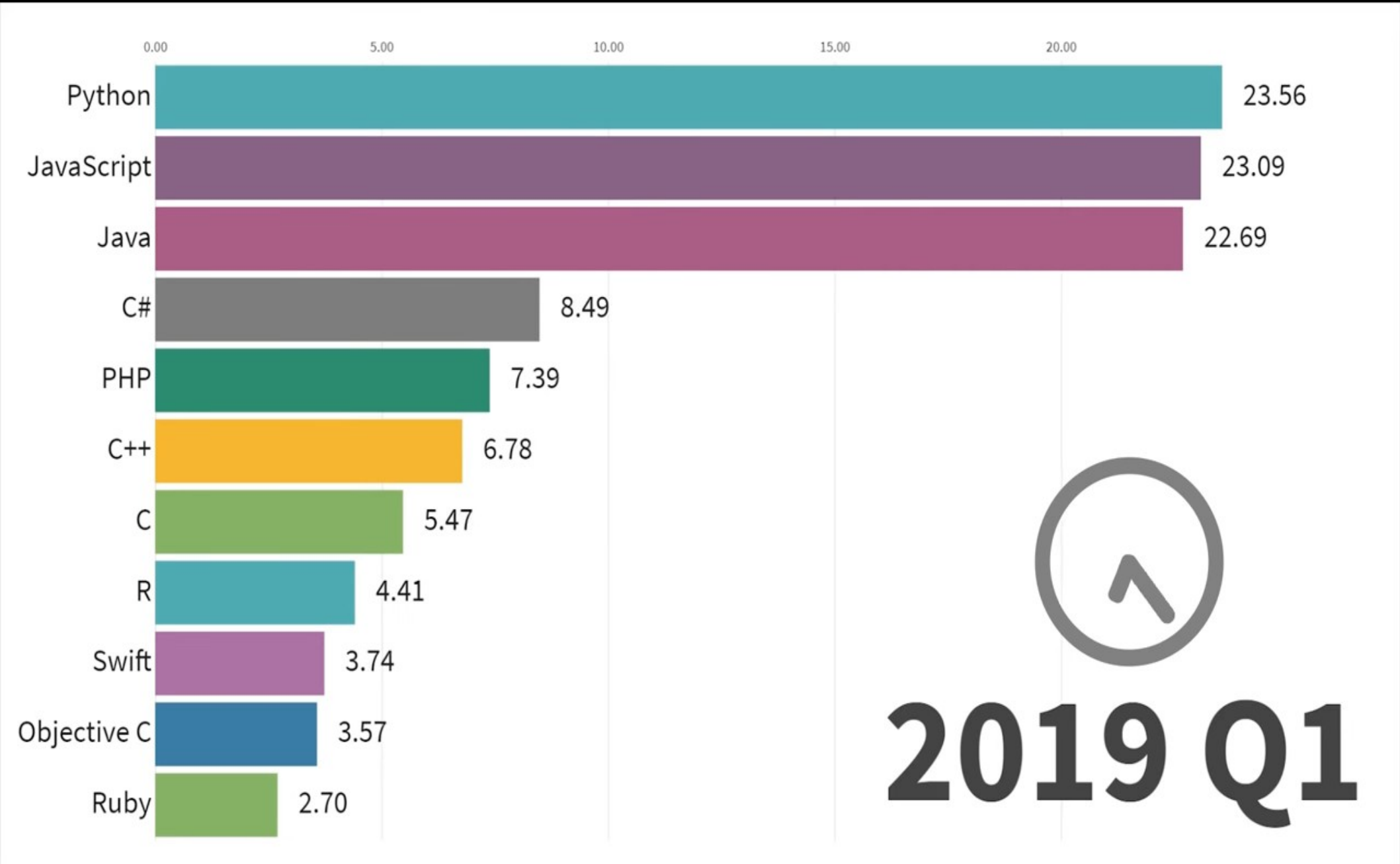
Most Popular Programming Languages 1965 - 2019



Most Popular Programming Languages 1965 - 2019



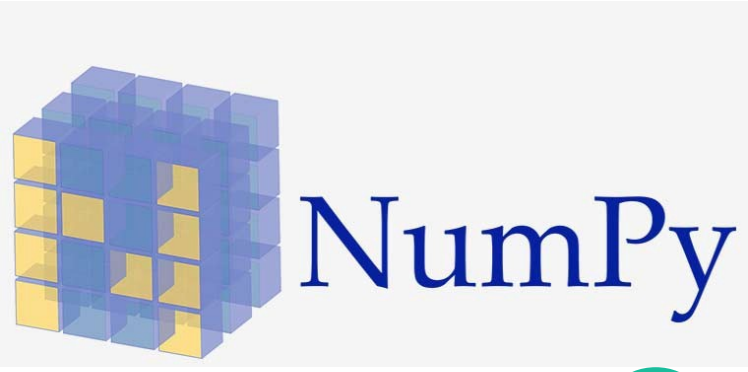
Most Popular Programming Languages 1965 - 2019



Bibliotecas Python

Bibliotecas: análise de dados - numpy

- Abreviatura de Numerical Python
- Ndarray: array multidimencional
- Funções matemáticas
- Álgebra linear



Bibliotecas: análise de dados - pandas

- Principal biblioteca exploração
- Leitura, manipulação e visualização
- Operação com séries, data frame
- I/O – csv, planilha eletrônica

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$

Bibliotecas: análise de dados - matplotlib

- **Biblioteca para a geração de gráficos 2D a partir de arrays**

matplotlib


ANACONDA.

Ananconda

Notebook

5.7.8

Web-based, interactive computing notebook environment. Edit and run human-readable docs while describing the data analysis.

Launch

Orange 3

3.19.0

Component based data mining framework. Data visualization and data analysis for novice and expert. Interactive workflows with a large toolbox.

Launch

Spyder

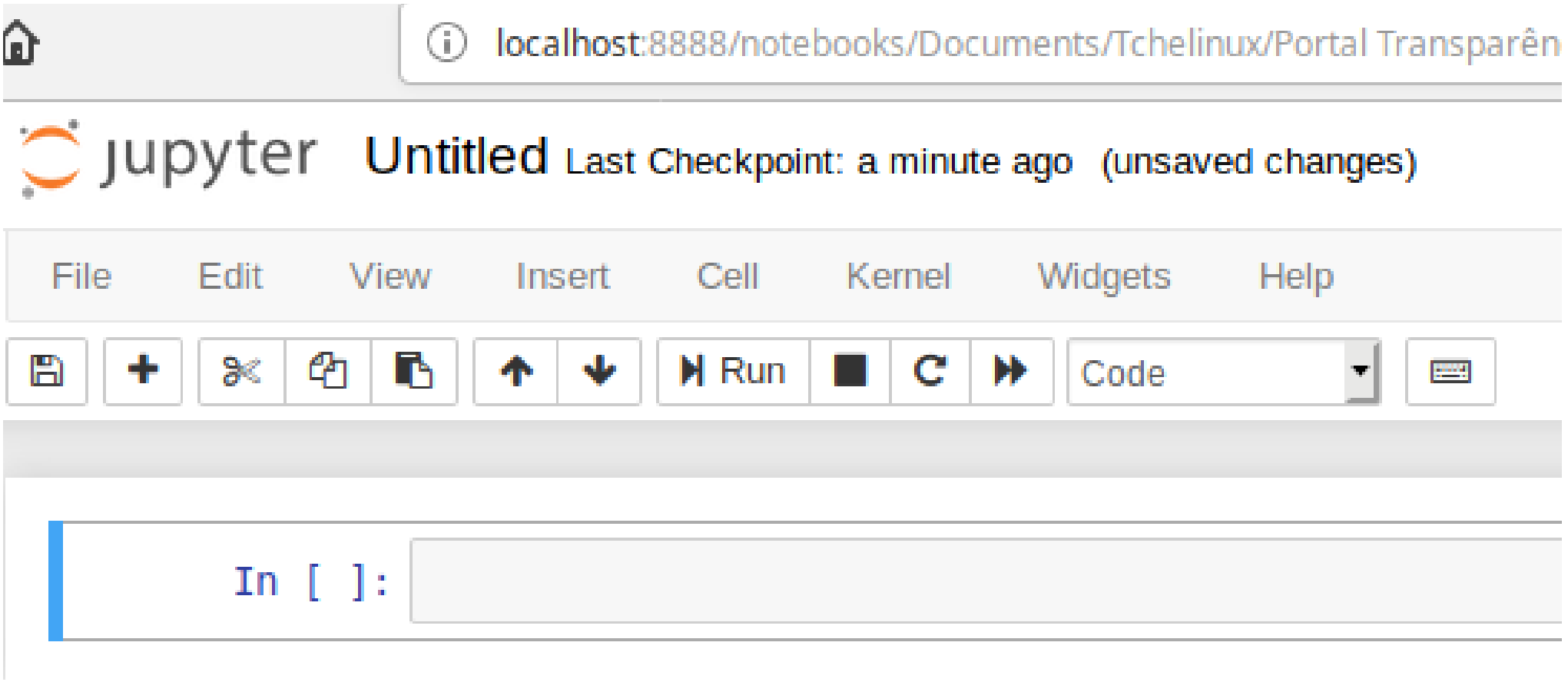
3.3.3

Scientific PYthon Development EnviRonment. Powerful Python IDE with advanced editing, interactive testing, debugging and introspection features

Launch


jupyter

Jupyter Notebook



Pandas

Pandas

- **Importação das Bibliotecas básicas**

```
import pandas as pd
import numpy as np
import matplotlib as plot
```

Pandas

- **Importar CSV e criar DataFrame**

```
data = pd.read_csv("dados.csv")
```

Pandas

- **Importar CSV e criar DataFrame**

```
data = pd.read_csv("dados.csv")
```

Pandas

- **Exibir dados do DataFrame**

  data

Exemplo de análise

Exibindo o conteúdo do arquivo

In [31]:
```
data
```

Out[31]:

Mes	Exercicio	FaseGasto	Favorecido	Poder	Cod_orgao	Orgao	Cod_UO	UO	Cod_Subfuncao	Subfuncao	
6.0	2019.0	Pagamento	FUNDO MUN DE SAUDE DE TRIUNFO	PODER EXECUTIVO	20.0	SECRETARIA DA SAUDE	2095.0	FUNDO ESTADUAL DE SAUDE	302	ASSIST.HOSP./AMBULATORIAL	5
6.0	2019.0	Pagamento	FUNDO MUN DE SAUDE DE JAGUARI	PODER EXECUTIVO	20.0	SECRETARIA DA SAUDE	2095.0	FUNDO ESTADUAL DE SAUDE	302	ASSIST.HOSP./AMBULATORIAL	5
6.0	2019.0	Pagamento	FUNDO MUN DE SAUDE DE JAGUARI	PODER EXECUTIVO	20.0	SECRETARIA DA SAUDE	2095.0	FUNDO ESTADUAL DE SAUDE	302	ASSIST.HOSP./AMBULATORIAL	5
6.0	2019.0	Pagamento	FUNDO MUN DE SAUDE DE MARCELINO	PODER EXECUTIVO	20.0	SECRETARIA DA SAUDE	2095.0	FUNDO ESTADUAL DE SAUDE	302	ASSIST.HOSP./AMBULATORIAL	5

Pandas

- **Exibir 5 primeiras linhas**

```
data.head()
```

Pandas

Mes	Exercicio	FaseGasto	Favorecido	Poder	Cod_orgao	Orgao	Cod_UO	UO	Cod_Subfuncao	Subfuncao	Cod_Acao
6.0	2019.0	Pagamento	FUNDO MUN DE SAUDE DE TRIUNFO	PODER EXECUTIVO	20.0	SECRETARIA DA SAUDE	2095.0	FUNDO ESTADUAL DE SAUDE	302	ASSIST.HOSP./AMBULATORIAL	562001003
6.0	2019.0	Pagamento	FUNDO MUN DE SAUDE DE JAGUARI	PODER EXECUTIVO	20.0	SECRETARIA DA SAUDE	2095.0	FUNDO ESTADUAL DE SAUDE	302	ASSIST.HOSP./AMBULATORIAL	562001003
6.0	2019.0	Pagamento	FUNDO MUN DE SAUDE DE JAGUARI	PODER EXECUTIVO	20.0	SECRETARIA DA SAUDE	2095.0	FUNDO ESTADUAL DE SAUDE	302	ASSIST.HOSP./AMBULATORIAL	562001003
6.0	2019.0	Pagamento	FUNDO MUN DE SAUDE DE MARCELINO RAMOS	PODER EXECUTIVO	20.0	SECRETARIA DA SAUDE	2095.0	FUNDO ESTADUAL DE SAUDE	302	ASSIST.HOSP./AMBULATORIAL	562001003
6.0	2019.0	Pagamento	FUNDO MUN DE SAUDE DE TAPES	PODER EXECUTIVO	20.0	SECRETARIA DA SAUDE	2095.0	FUNDO ESTADUAL DE SAUDE	302	ASSIST.HOSP./AMBULATORIAL	562001003

Pandas

- **Filtrar dados**
  - loc: baseado labels das colunas
    array booleano

data.loc[<linhas>, <colunas>]

Pandas - loc

data.loc[<linhas>, <colunas>]

data.loc[5]

data.loc[3:7]

data.loc[[0,1,2]]

Pandas - loc

```
data.loc[(data["Favorecido"]==
"FUNDO MUN DE SAUDE DE
ALEGRETE")]
```

Exemplo de análise

```
data.loc[(data["Favorecido"]=="FUNDO MUN DE SAUDE DE ALEGRETE")]
```

Mes	Exercicio	FaseGasto	Favorecido	Poder	Cod_orgao	Orgao	Cod_UO	UO
6.0	2019.0	Pagamento	FUNDO MUN DE SAUDE DE ALEGRETE	PODER EXECUTIVO	20.0	SECRETARIA DA SAUDE	2095.0	FUNDO ESTADUAL DE SAUDE
6.0	2019.0	Pagamento	FUNDO MUN DE SAUDE DE ALEGRETE	PODER EXECUTIVO	20.0	SECRETARIA DA SAUDE	2095.0	FUNDO ESTADUAL DE SAUDE

Pandas

- **Filtrar dados**
  - iloc: números inteiros linhas

data.loc[<linhas>, <colunas>]

Pandas - iloc

```
data.iloc[0]    # 1ª linha do dataset

data.iloc[-1]   # última linha
```

Pandas - iloc

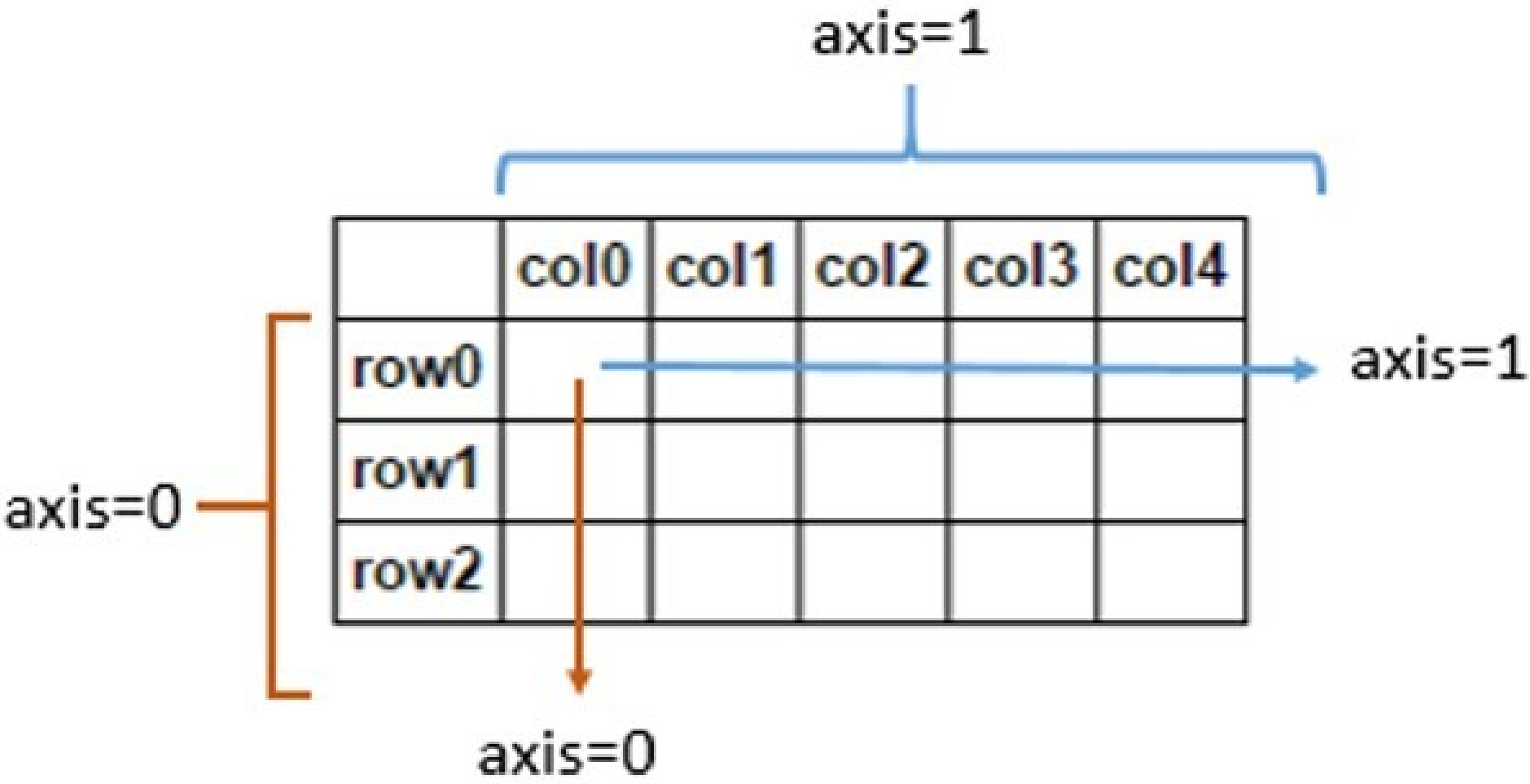
```
data.iloc[:,0]

# Todos os dados da 1ª coluna


data.iloc[0:5,-1]

# 1º ao 5º dado da última coluna
```

Pandas

- **Dropando colunas**

```
data = data.drop('NomCol', axis=1)
```

Pandas


axis=1
col0
col1
col2
col3
col4
row0
row1
row2
axis=1
axis=0
axis=0

Pandas

- **Dropando múltiplas colunas**

```
data.drop(['Col1', 'Col3','Col2'],
axis=1)
```

Pandas

- **Dropando múltiplas colunas**

```
for col in data.columns:
    if col not in ('Col3','Col5','Col7')
        del data[col]
```

Pandas

- **Dropando 2 últimas linhas**

```
data.drop(data.tail(2).index)
```

Pandas

- **Somando valores**

```
data["NewCol"] =
                data['Col1'] + data['Col2']
```

Pandas

- **Somatório Groupby**

```
data.groupby(
        ['Col1','Col2'], as_index=False
    )['ColValue1','ColValue2'].sum()
```

Pandas

- **Export csv**

```
data.to_csv("output.csv", index =
None, header=True)
```

Perguntas ?

Obrigado !!!

andriusjaques@pm.me

github.com/andriusjaques

Links e referências

- https://www.anaconda.com
- https://jupyter.org/
- https://pandas.pydata.org/
- https://panda.ime.usp.br/algoritmos/static/algoritmos/10-matplotlib.html
- https://panda.ime.usp.br/algoritmos/static/algoritmos/10-matplotlib.html
- https://medium.com/horadecodar/data-science-tips-02-como-usar-loc-e-iloc-no-pandas-fab58e214d87
- https://www.vooo.pro/insights/guia-de-acesso-rapido-ao-pandas/
- https://www.alura.com.br/artigos/criando-graficos-no-python-com-a-matplotlib
- https://paulovasconcellos.com.br/15-comandos-de-matplotlib-que-talvez-voc%C3%AA-n%C3%A3o-conhe%C3%A7a-17cf88a75119
- https://www.it-swarm.net/pt/python/qual-e-o-significado-do-atributo-axis-em-um-dataframe-pandas/826546083/
- https://www.logiquesistemas.com.br/blog/ciencia-de-dados/