

MVP - Engenharia de Dados

PUC-RIO.

Aluno: Luís Fernando Andrade Cordeiro

Objetivo

O presente estudo tem como objetivo principal analisar a evolução da população total ao longo de um período específico, destacando a posição do Brasil em relação a outros países.

Detalhamento

Busca pelos dados

<https://www.kaggle.com/datasets/alitaqi000/global-population-trends2016-2022>

 SYED ALI TAQI · ATUALIZADO HÁ UM MÊS

34

Novo caderno

Download icon

Baixar (32kB)

Tendências da População Global (2016-2022)

Este conjunto de dados contém informações demográficas de quase 215 países (2016 a 2022)



Cartão de dados

Código (4)

Discussão (1)

Sobre o conjunto de dados

Conjunto de dados de tendências populacionais globais:

explore uma coleção abrangente dos principais indicadores demográficos de todo o mundo. Este conjunto de dados abrange a população total, distribuições urbano-rurais, esperança de vida, taxas de natalidade e mortalidade, taxa de fertilidade, taxa de mortalidade infantil e taxa de crescimento. Descubra insights sobre a dinâmica populacional global e seus fatores subjacentes. Ideal para pesquisas, visualizações e análises.

Conteúdo:

- 1) População total
- 2) População urbana e rural
- 3) Densidade populacional (a densidade populacional é a população no meio do ano dividida pela área terrestre em quilômetros quadrados)
- 4) Expectativa de vida (a expectativa de vida ao nascer indica o número de anos que um recém-nascido viveria se os padrões de mortalidade prevalentes no momento de seu nascimento permanecessem os mesmos ao longo de sua vida) 5) Taxas de natalidade e mortalidade (taxa bruta indica

a número de mortes ocorridas durante o ano, por 1.000 habitantes estimado em meados do ano)

6) Taxa de fertilidade (a taxa de fertilidade total representa o número de filhos que nasceriam de uma mulher se ela vivesse até o fim de sua

Usabilidade ⓘ

10h00

Licença

Banco de dados: Banco de dad...

Frequência de atualização esperada

Anualmente

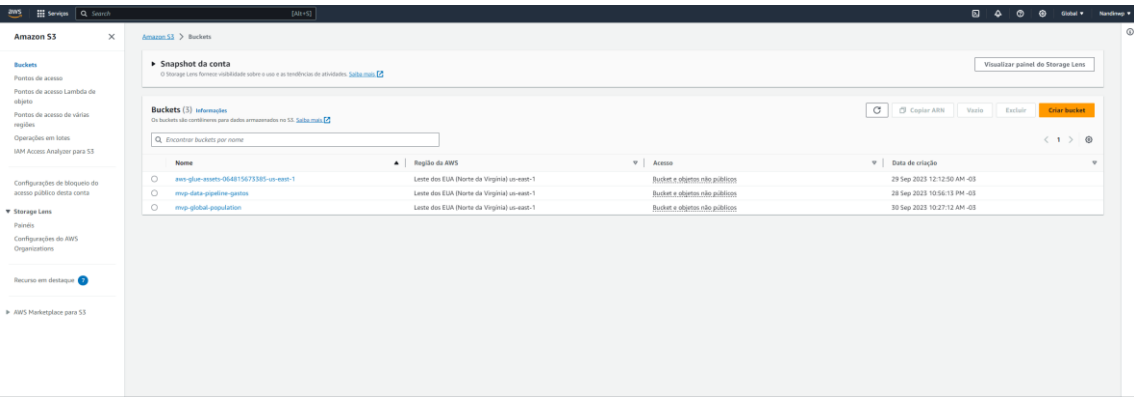
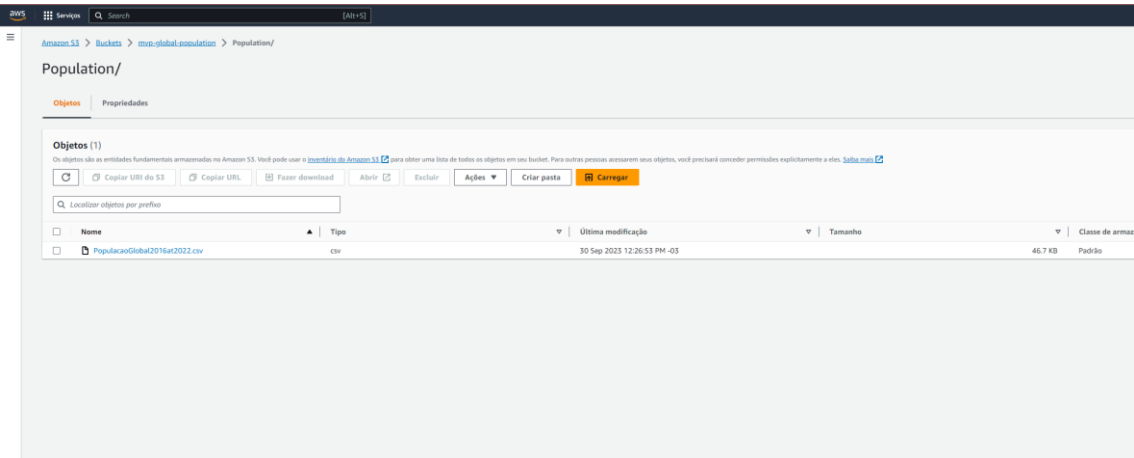
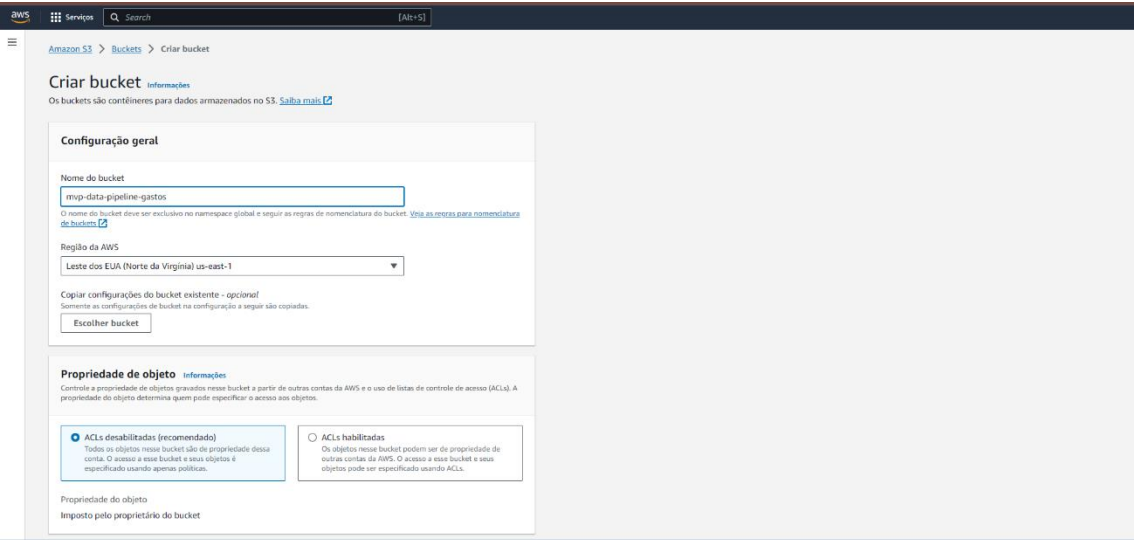
Tag

Ciências Sociais

Coleta

A primeira etapa consistiu na realização do download dos dados diretamente do repositório Kaggle, uma plataforma amplamente reconhecida pela disponibilização de conjuntos de dados em diversas áreas de estudo. Após o download, os dados foram verificados e preparados para a próxima fase do processo.

O próximo passo envolveu a inserção manual dos dados preparados no Amazon S3, um serviço de armazenamento em nuvem altamente escalável e seguro disponibilizado pela AWS. Este procedimento foi realizado por meio da interface de usuário fornecida pela plataforma, onde os arquivos foram devidamente organizados em diretórios relevantes para facilitar a posterior manipulação e análise.



Modelagem

GlobalPp	
Country 	varchar
Year	varchar
TotalPopulation	varchar
UrbanPopulation	varchar
RuralPopulation	varchar
PopulationDensity	varchar
LifeExpectancy	varchar
BirthRate	varchar
DeathRate	varchar
FertilityRate	varchar
InfantMortalityRate	varchar
GrowthRate	varchar

Legenda:

Country – contém os países (100 no total)

Year – Ano da coleta das informações

TotalPopulation – contém a quantidade total da população

UrbanPopulation - contém a quantidade da população urbana

RuralPopulation – contém a quantidade da população Rural

PopulationDensity – Densidade da população (ex: 20 habitantes por km²)

LifeExpectancy – Expectativa de vida

BirthRate – Taxa de natalidade por mil habitantes

DeathRate – taxa de mortalidade por mil habitantes

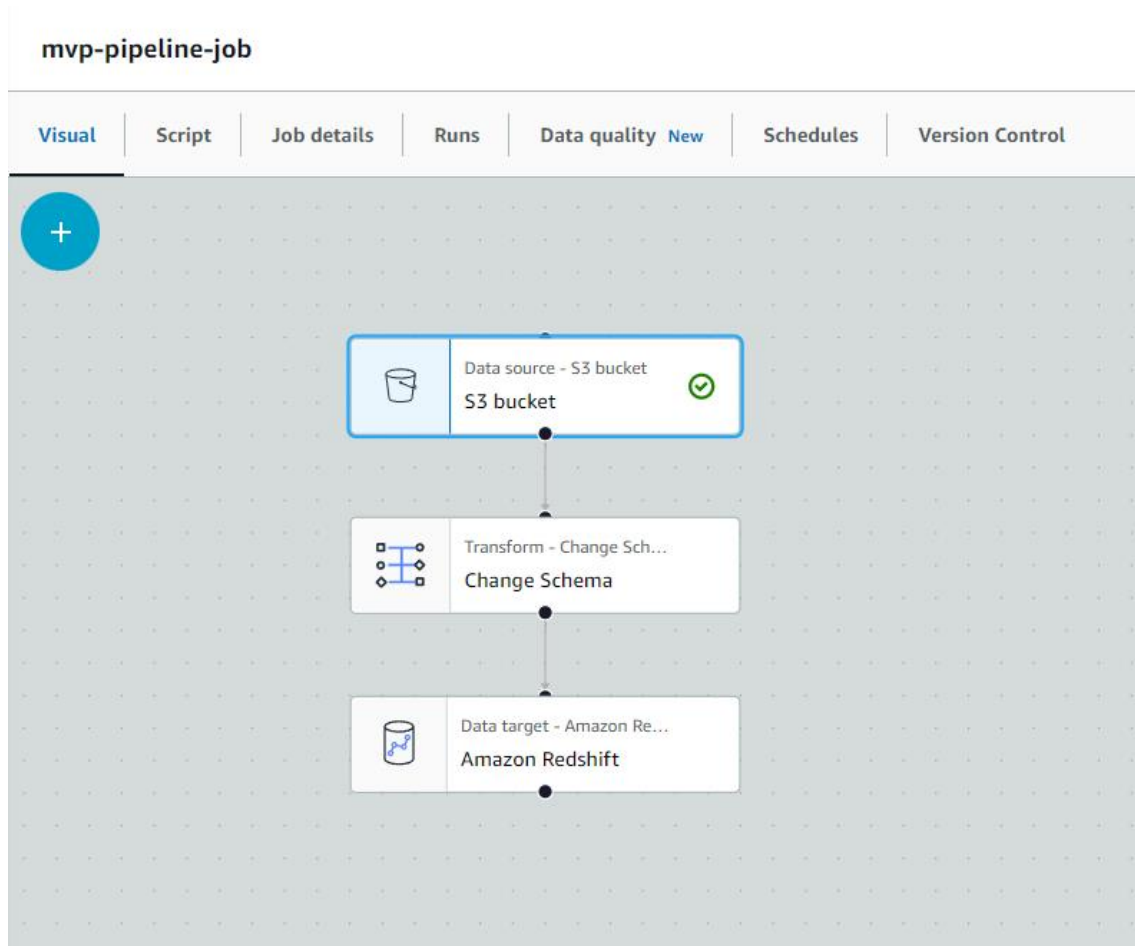
FertilityRate – taxa de fertilidade

InfantMortalityRate – taxa de mortalidade infantil

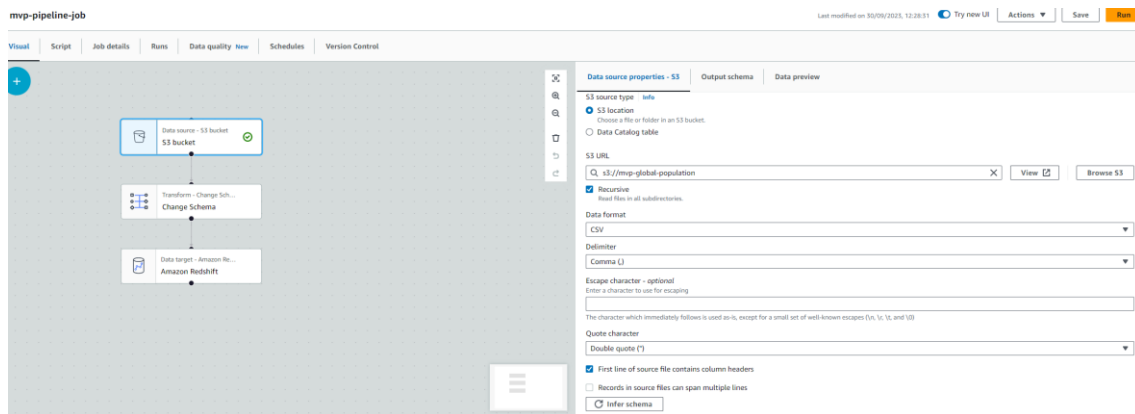
GrowthRate - taxa de crescimento

Carga

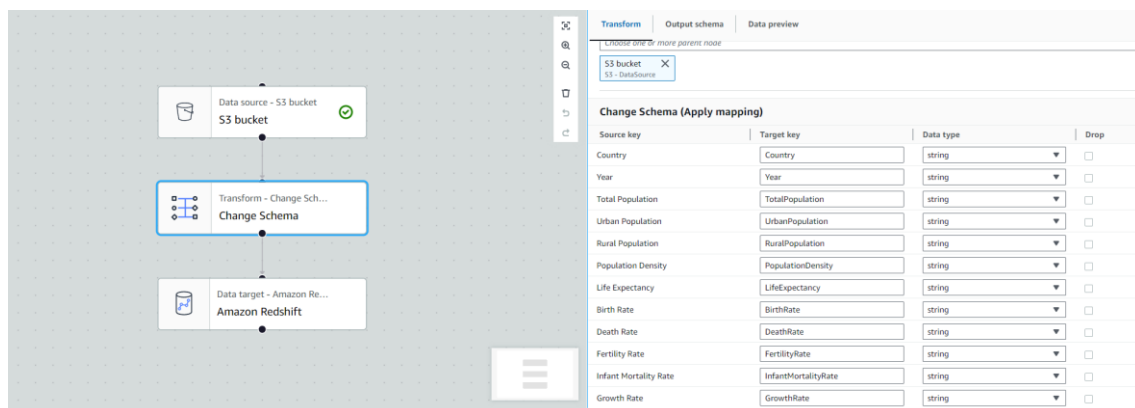
O processo de ETL (Extração, Transformação e Carga) foi conduzido por meio do serviço AWS Glue, uma plataforma oferecida pela Amazon Web Services dedicada à execução de tarefas de integração de dados de forma eficiente e escalável. Por meio da interface gráfica proporcionada pelo AWS Glue, foram delineadas e implementadas as seguintes fases:



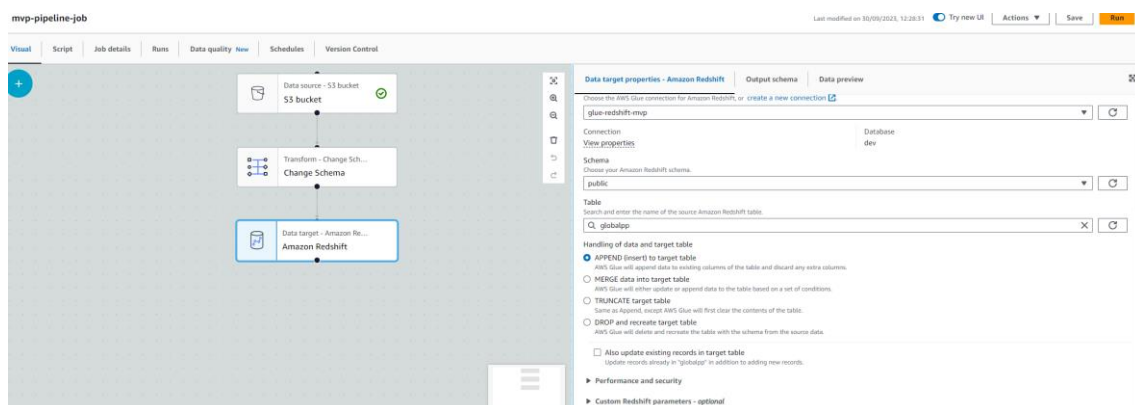
1. Extração: Nesta etapa, os dados foram obtidos de suas fontes originais. O AWS Glue facilitou esse processo ao permitir a conexão com diversas fontes de dados, como bancos de dados, serviços de armazenamento e APIs. Dessa forma, os conjuntos de dados foram identificados e selecionados para posterior transformação. Dados do meu S3 “mvp-global-population” na pasta “Population”.



2. Transformação (Transform – Change Schema): A fase de transformação é fundamental para preparar os dados para análise. Por meio do AWS Glue, foram aplicadas uma série de operações, como limpeza, agregação, filtragem e conversão de formatos. Além disso, foram realizadas manipulações específicas para adequar os dados ao modelo de análise, neste caso optei em deixar todos os campos originais, removendo apenas os espaços das colunas deixando no padrão camelcase.



3. Carga: Após a transformação - O AWS Glue permitiu configurar esse processo de forma automatizada e escalável, garantindo a eficiência na persistência dos dados.



E por fim executei o job para a que todas as configurações sejam registradas.

mvp-pipeline-job

Last modified on 30/09/2023, 12:28:31Try new UIActions

VisualScriptJob detailsRunsData qualityNewSchedulesVersion Control

Job runs (1/8)Info

Last updated (UTC)September 30, 2023 at 15:29:28View detailsStop job runTable View

Filter job runs by property

Run status	Retries	Start time	End time	Duration	Capacity (DPUs)	Worker type	Glue version
Running	0	09/30/2023 12:28:33	-	26 s	2 DPUs	G.1X	4.0
Succeeded	0	09/30/2023 11:16:58	09/30/2023 11:20:08	2 m 35 s	2 DPUs	G.1X	4.0
Succeeded	0	09/30/2023 11:07:55	09/30/2023 11:11:05	2 m 16 s	2 DPUs	G.1X	4.0
Succeeded	0	09/30/2023 10:05:24	09/30/2023 10:08:25	2 m 33 s	2 DPUs	G.1X	4.0
Succeeded	0	09/30/2023 09:26:53	09/30/2023 09:29:59	2 m 34 s	2 DPUs	G.1X	4.0
Failed	0	09/29/2023 00:28:39	09/29/2023 00:31:50	2 m 43 s	2 DPUs	G.1X	4.0
Failed	0	09/29/2023 00:20:37	09/29/2023 00:21:39	57 s	2 DPUs	G.1X	4.0
Failed	0	09/29/2023 00:17:07	09/29/2023 00:20:05	2 m 25 s	10 DPUs	G.1X	4.0

09/30/2023 11:16:58

Job nameIdRun statusGlue version

mvp-pipeline-jobjr_023f2275189ef7d94375a368ea30e358b2b43376f0b97a03288e2361e76f23407Succeeded4.0

Retry attempt numberStart timeEnd timeStart-up time

Initial run30 de setembro de 2023 11:16:5830 de setembro de 2023 11:20:0834 seconds

Execution timeLast modified onTrigger nameSecurity configuration

Análise

A – Qualidade dos dados

Os dados obtidos no Kaggle apresentaram poucos valores faltantes onde avia “-“no lugar da informação, para corrigir este problema antes mesmo do upload dos arquivos locais para o S3, fiz um pequeno processamento deste dados que resultou em sua remoção por completo.

import pandas as pd

Carrega o CSV

df = pd.read_csv('/content/PopulacaoGlobal2016at2022.csv', encoding='utf-8')

Remove linhas que contém "-"

df = df[(df == '-').any(axis=1)]

Remove a vírgula de Total Population, Urban Population e Rural Population

df['Total Population'] = df['Total Population'].str.replace(',', '')

df['Urban Population'] = df['Urban Population'].str.replace(',', '')

df['Rural Population'] = df['Rural Population'].str.replace(',', '')

Salva o Dataframe de volta no CSV

df.to_csv('/content/PopulacaoGlobal2016at2022NEW.csv', index=False, encoding='utf-8')

df.head()

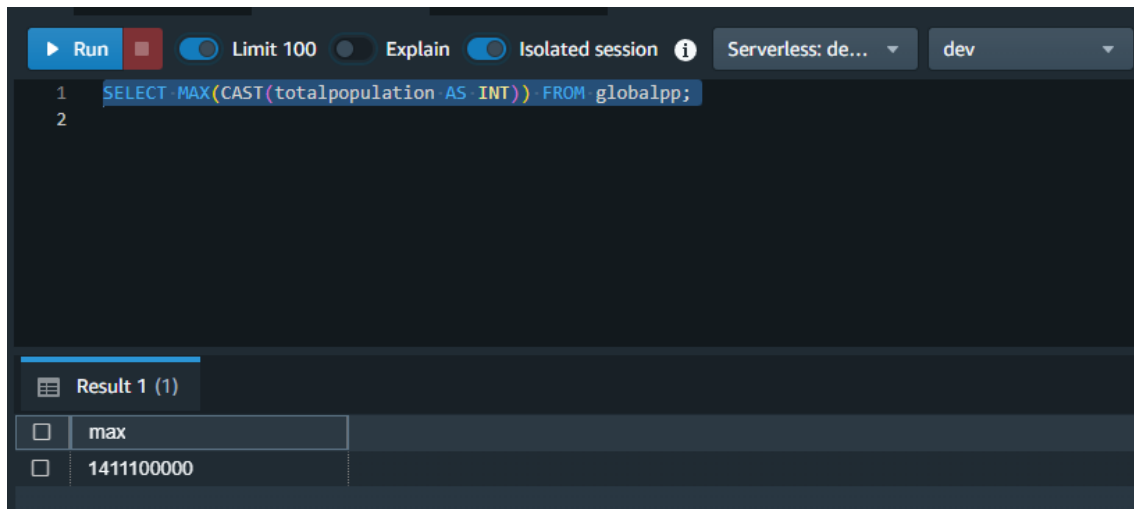
	Country	Year	Total Population	Urban Population	Rural Population	Population Density	Life Expectancy	Birth Rate	Death Rate	Fertility Rate	Infant Mortality Rate	Growth Rate
0	Afghanistan	2018	36686784	9353296	27333488	56	63.0	36.927	6.981	5.002	47.8	3.0
1	Afghanistan	2019	37769499	9727157	28042342	58	64.0	36.466	6.791	4.870	46.3	3.0
2	Afghanistan	2020	38972230	10142913	28829317	60	63.0	36.051	7.113	4.750	44.8	3.0
3	Albania	2018	2866376	1728969	1137407	105	79.0	10.517	8.308	1.440	8.3	0.0
4	Albania	2019	2854191	1747593	1106598	104	79.0	10.343	8.480	1.414	8.4	0.0

B – Solução do problema

B.1 Primeiramente vamos consultar algumas informações para podermos trabalhar em cima delas.

Partimos então para a consulta: `SELECT MAX(CAST(totalpopulation AS INT)) FROM globalpp;`

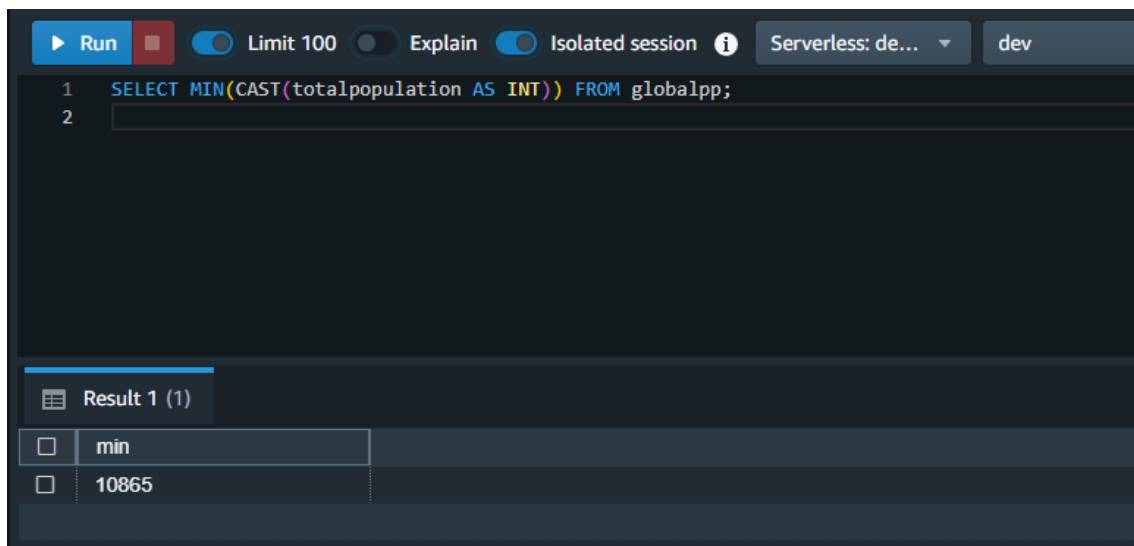
Que me retorna o valor máximo desta coluna, neste caso estou usando cast as int pois defini no essas tabela como string



The screenshot shows a SQL query editor with a dark theme. At the top, there are buttons for 'Run', 'Limit 100', 'Explain', 'Isolated session', and a dropdown menu for 'Serverless: de...' with 'dev' selected. The query text is: `1 SELECT MAX(CAST(totalpopulation AS INT)) FROM globalpp;`
`2`
Below the query, the results are displayed under the heading 'Result 1 (1)'. The results table has two rows: the first row has a header 'max' and the second row has the value '1411100000'.

max
1411100000

Faremos novamente, só que agora usando **MIN** para obter o menor valor.



The screenshot shows the same SQL query editor as before. The query text is: `1 SELECT MIN(CAST(totalpopulation AS INT)) FROM globalpp;`
`2`
Below the query, the results are displayed under the heading 'Result 1 (1)'. The results table has two rows: the first row has a header 'min' and the second row has the value '10865'.

min
10865

Agora vamos ver quais são os 5 países com a maior população:

Neste caso estou realizando uma consulta usando SUM para somar com as tabelas que se repetem, por exemplo: se China aparece 3 vezes pois temos dados de 3 anos, esses valores vão ser somados, por fim eu ordeno todos do maior para o menor:

```

1  SELECT country, SUM(CAST(totalpopulation AS INT)) AS populacao_total
2  FROM GlobalPp
3  GROUP BY country
4  ORDER BY populacao_total DESC
5  LIMIT 5;
6

```

Result 1 (5)

<input type="checkbox"/>	country	populacao_total	
<input type="checkbox"/>	China	4221605000	
<input type="checkbox"/>	India	4148502483	
<input type="checkbox"/>	United States	986679664	
<input type="checkbox"/>	Indonesia	808507691	
<input type="checkbox"/>	Pakistan	670221500	


Como podemos analisar, a China é o mais populoso. Porém, a Índia tem uma diferença bem pequena em comparação com a China. Cerca de 1,73%. Seguindo, temos Estados Unidos, Indonésia e Pakistan.

Agora, onde fica a posição do Brasil em relação a estes que foram analisados? Simples, ele fica em 6º lugar.


```

1 SELECT country, SUM(CAST(totalpopulation AS INT)) AS populacao_total
2 FROM GlobalPp
3 GROUP BY country
4 ORDER BY populacao_total DESC
5 LIMIT 10
6

```

 Result 1 (10)

<input type="checkbox"/>	country	populacao_total	
<input type="checkbox"/>	China	4221605000	
<input type="checkbox"/>	India	4148502483	
<input type="checkbox"/>	United States	986679664	
<input type="checkbox"/>	Indonesia	808507691	
<input type="checkbox"/>	Pakistan	670221500	
<input type="checkbox"/>	Brazil	635145774	
<input type="checkbox"/>	Nigeria	610019520	
<input type="checkbox"/>	Bangladesh	496621131	
<input type="checkbox"/>	Russia	432957259	
<input type="checkbox"/>	Japan	379705000	

Certo, mais ainda não ficou muito claro e nenhum pouco intuitivo, vamos agora mostrar essas informações um pouco mais detalhada:

```
import matplotlib.pyplot as plt

países = ['China', 'India', 'United States', 'Indonesia', 'Pakistan', 'Brazil', 'Nigeria', 'Bangladesh', 'Russia', 'Japan']
populacoes = [4221605000, 4148502483, 986679664, 808507691, 670221500, 635145774, 610019520, 496621131, 432957259, 379705000]

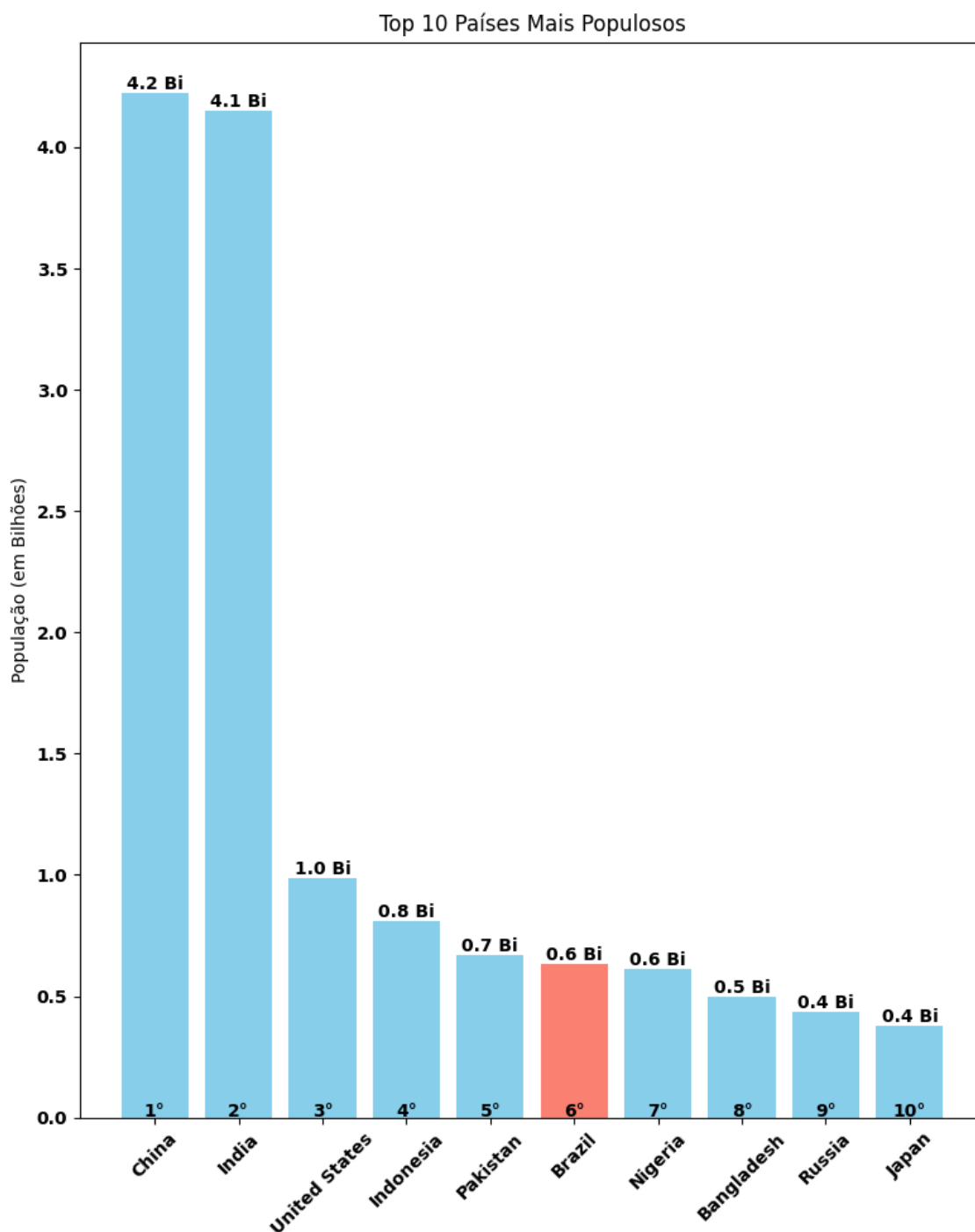
# Converte os valores para bilhões
populacoes_bi = [pop/1e9 for pop in populacoes]

plt.figure(figsize=(8, 10))
bars = plt.bar(países, populacoes_bi, color=['skyblue']*5 + ['salmon'] + ['skyblue']*4) # Destaca o Brasil em vermelho
plt.ylabel('População (em Bilhões)')
plt.title('Top 10 Países Mais Populosos')

# Adiciona os valores nas barras
for i, bar in enumerate(bars):
    height = bar.get_height()
    plt.text(bar.get_x() + bar.get_width()/2., height, f'{height:.1f} Bi', ha='center', va='bottom', fontsize=10, fontweight='bold')
    plt.text(bar.get_x() + bar.get_width()/2., -0.01, f'{i+1}º', ha='center', va='bottom', fontsize=10, fontweight='bold')

plt.xticks(rotation=45) # Rotaciona os nomes dos países
plt.tight_layout()

plt.show()
```



Conclusão:

Como visto nesta análise, o Brasil é o 6° país mais populoso, perdendo do Pakistan e ficando a frente da Nigeria. Já a China domina em primeiro lugar com uma diferença de 1.73% em relação a Índia.

O presente estudo se propõe a examinar primariamente a evolução da população total ao longo de um período específico, com o intuito de destacar a posição relativa do Brasil em comparação com outras nações.

A solução adotada para lidar com valores ausentes, representados pelo símbolo "-", demonstrou um procedimento metodológico meticuloso e efetivo. A etapa inicial de pré-processamento dos dados localmente, anterior ao carregamento no serviço S3, revela uma compreensão perspicaz das exigências do projeto.

A utilização do operador de conversão **CAST** para transformar o tipo de dado da coluna "totalpopulation" em inteiro foi uma abordagem apropriada para viabilizar a ordenação precisa dos países com base em sua população. Esta escolha indica um entendimento aprofundado das particularidades dos dados e do SQL.

A identificação dos cinco países mais populosos e a análise da discrepância percentual entre a China e a Índia foram conduzidas de forma clara e concisa. A inclusão da posição do Brasil na lista adiciona um contexto valioso, permitindo uma apreensão mais abrangente do cenário demográfico global.

A discussão sobre a disparidade percentual entre a China e a Índia realça a proximidade demográfica entre essas duas nações, o que pode ter implicações significativas em diversos âmbitos, como economia, políticas públicas e desafios sociais.

Por fim, a conclusão ressalta a importância do Brasil como o sexto país mais populoso e contextualiza sua posição em relação a outras nações objeto de análise. Ademais, a comparação entre China e Índia enfatiza a magnitude das populações desses países e suas ramificações globais.

De maneira abrangente, a solução apresentada reflete um entendimento sólido dos dados e uma abordagem analítica eficaz para explorar e interpretar as tendências populacionais globais.

Autoavaliação

Com base na análise dos dados apresentados, conclui-se que foi possível atender às indagações propostas, valendo-se da base de dados que dispõe de informações acerca da população global. É importante ressaltar que essa conclusão está pautada no contexto vigente em 30 de setembro de 2023, data em que se observa a Índia ultrapassando a China, tornando-se o país mais populoso.

Ao longo do processo, surgiu a complexidade na elaboração de uma documentação mais robusta, destacando-se a dificuldade em estabelecer a conexão entre o banco de dados e as plataformas do Redshift e Jupyter

Notebook. Não obstante, os resultados obtidos são consideráveis e fornecem uma base sólida para futuras análises, nas quais pretendo explorar com mais profundidade os dados disponíveis nas tabelas disponibilizadas.