MVP – ENGENHARIA DE DADOS

Aluno: Matheus de Almeida Folly Ribeiro

Análise da expectativa de vida em 1990 a 2021

VISÃO GERAL

Objetivos do projeto

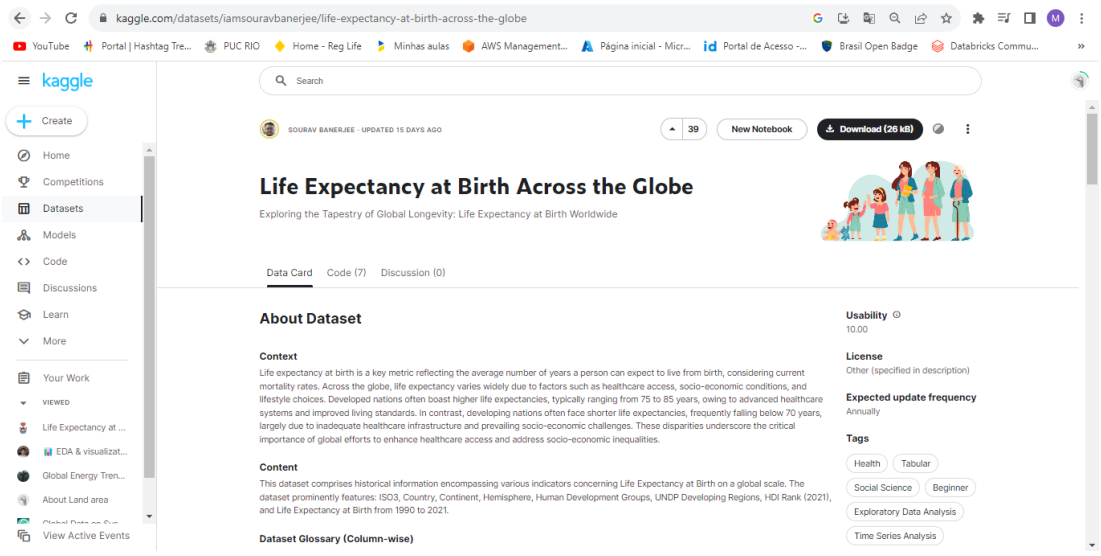
O projeto tem como objetivo analisar a evolução da expectativa de vida ao redor do mundo. A análise terá como foco:

- Qual a média de expectativa de vida por continente em 2021?
- Quais são os 10 países com melhores posições no IDH 2021? A quais continentes pertencem?
- Qual a correlação entre a expectativa de vida e o ranking do IDH em 2021?
- Quais países tiveram as maiores variações positivas na expectativa de vida entre 1990 a 2021? Algum desses países está entre os 10 primeiros colocados no ranking do IDH 2021?

DETALHAMENTO

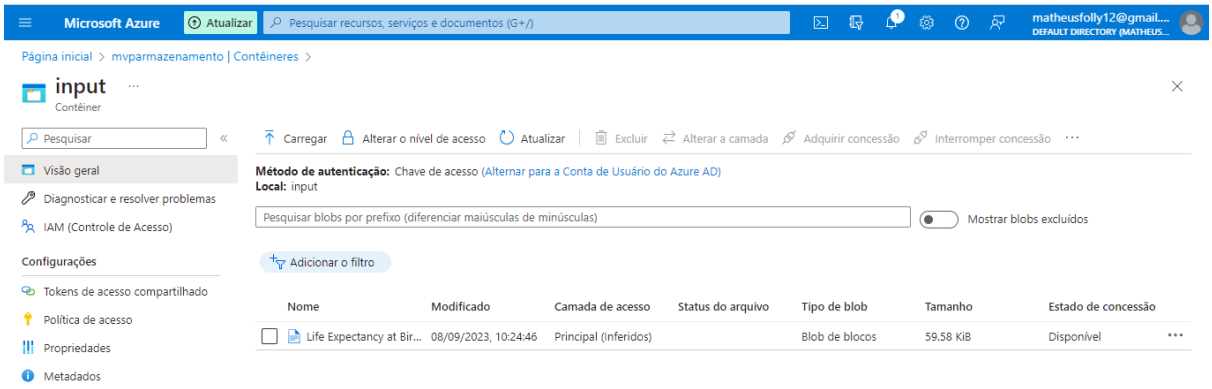
1. Busca pelos dados

O dataset foi retirado do KAGGLE:



2. Coleta

Realizei o download do arquivo em formato csv para minha máquina local e fiz o upload para um contêiner na Microsoft Azure.



3. Modelagem

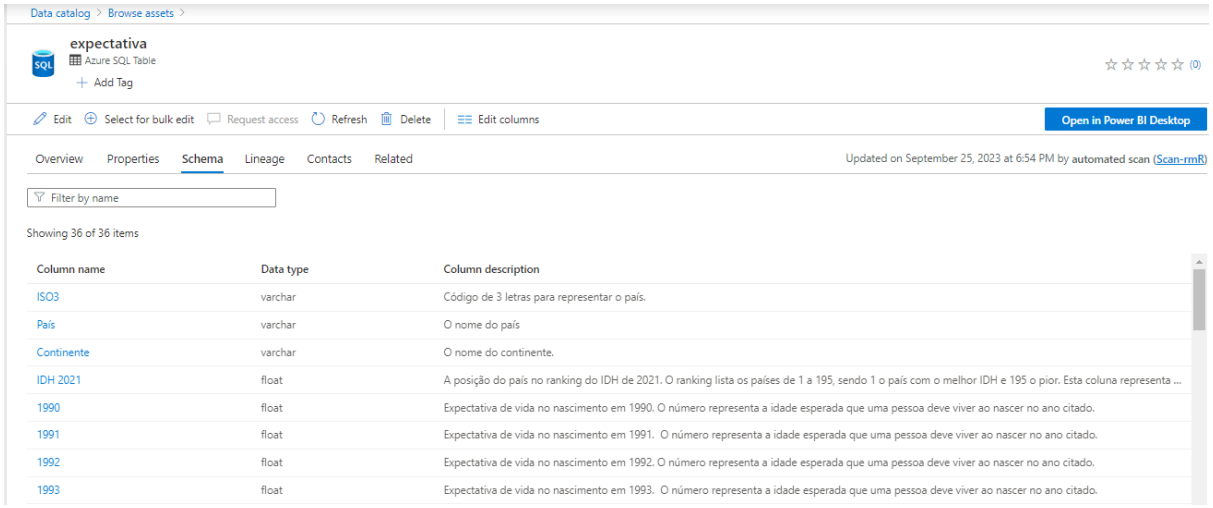
Como modelo de dados, foi utilizado uma tabela flat.

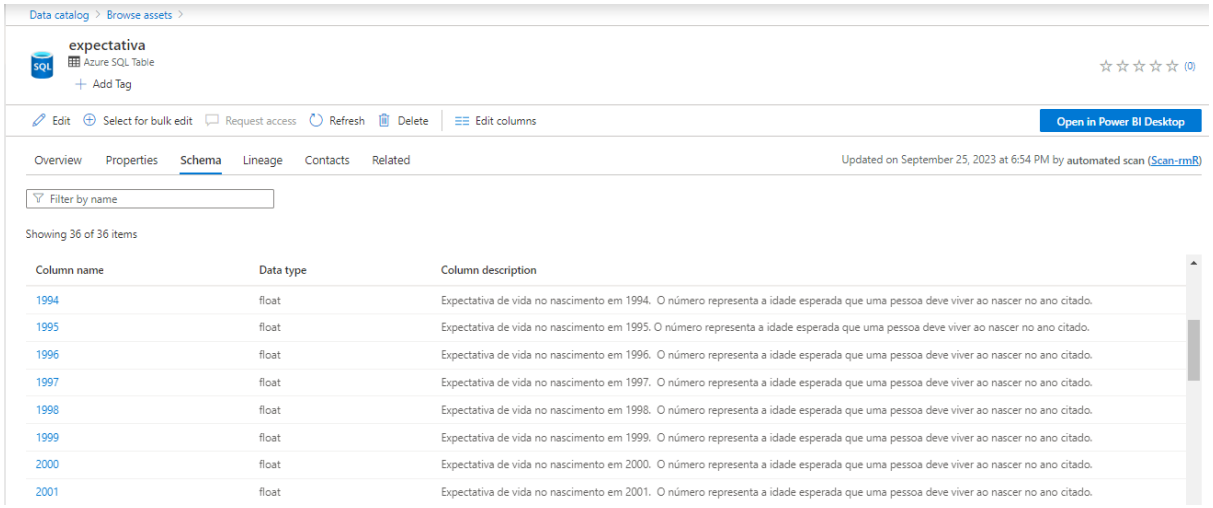
A base de dados foi retirada do Kaggle: https://www.kaggle.com/datasets/iamsouravbanerjee/life-expectancy-at-birth-across-the-globe

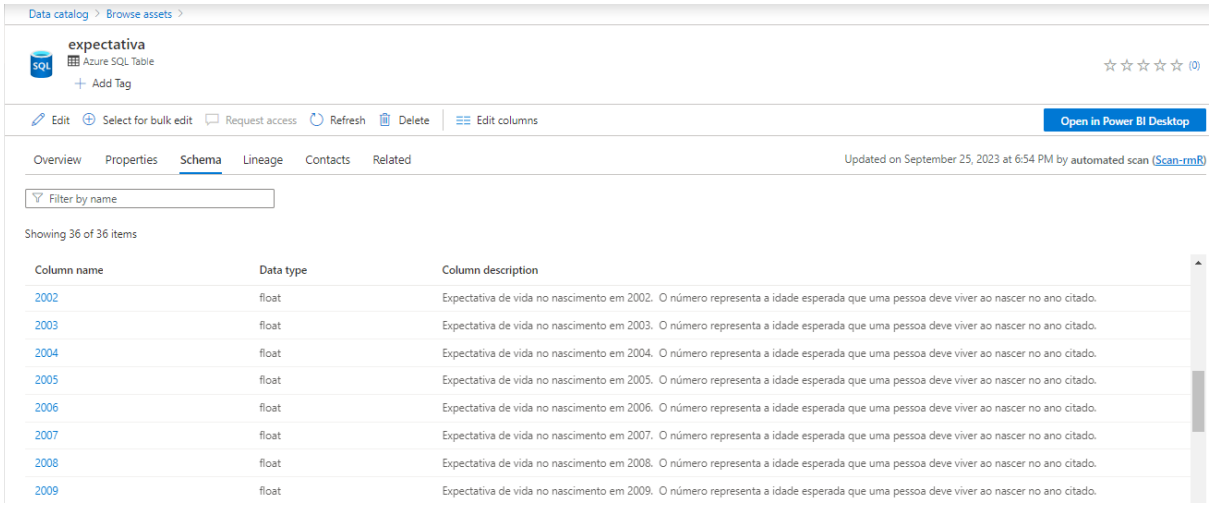
O autor do dataset no Kaggle retirou os dados do relatório de desenvolvimento humano feito pelo PNDU (Programa das Nações Unidas para o Desenvolvimento – ONU): https://www.undp.org.

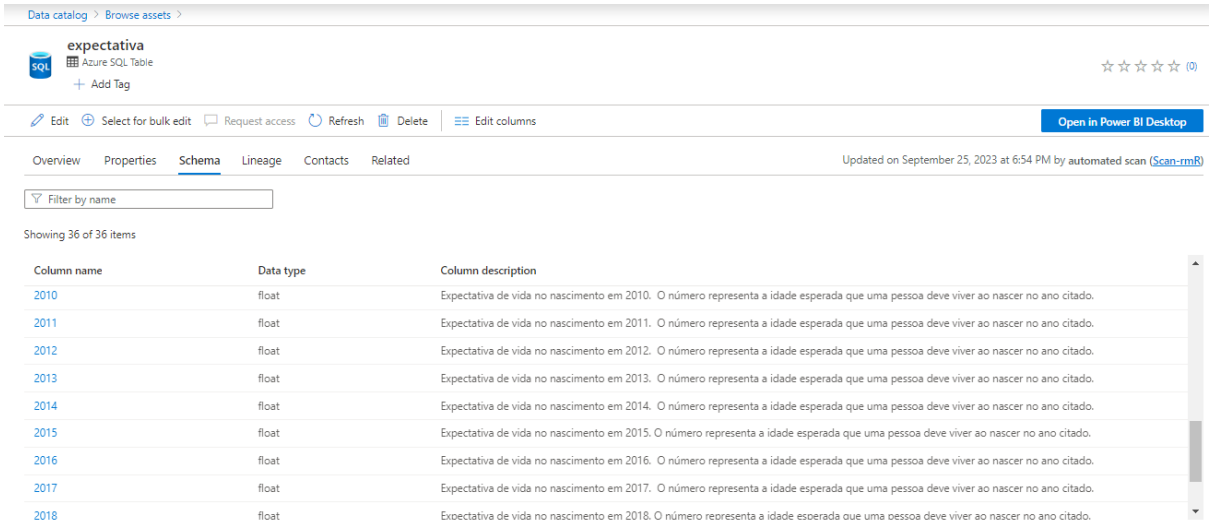
O arquivo foi baixado em CSV e foi feito o upload na plataforma da Microsoft Azure para o tratamento e armazenamento dos dados em nuvem.

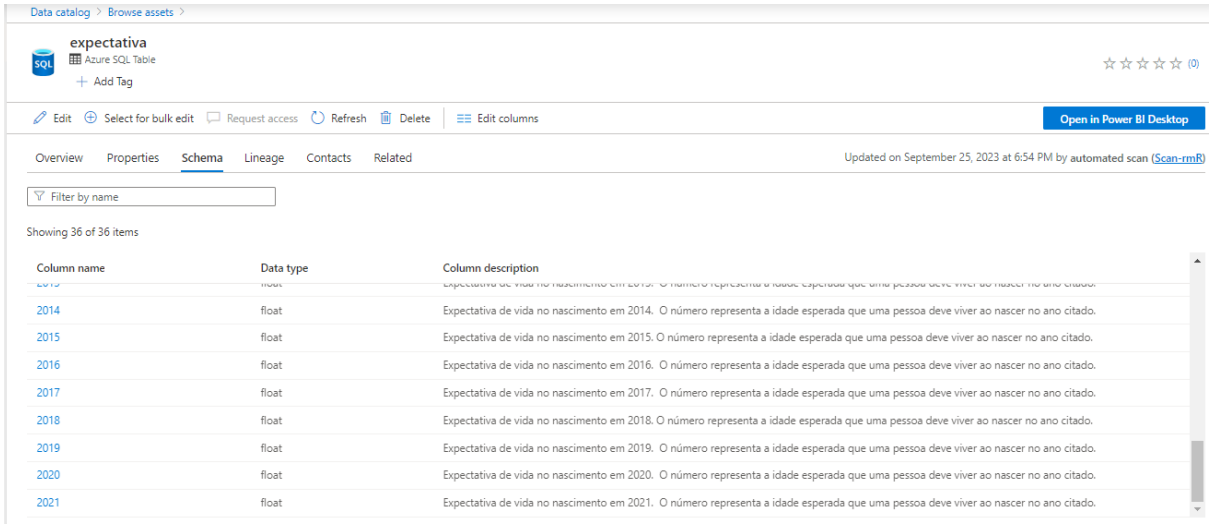
O Catálogo de dados foi feito utilizando o Microsoft Purview:









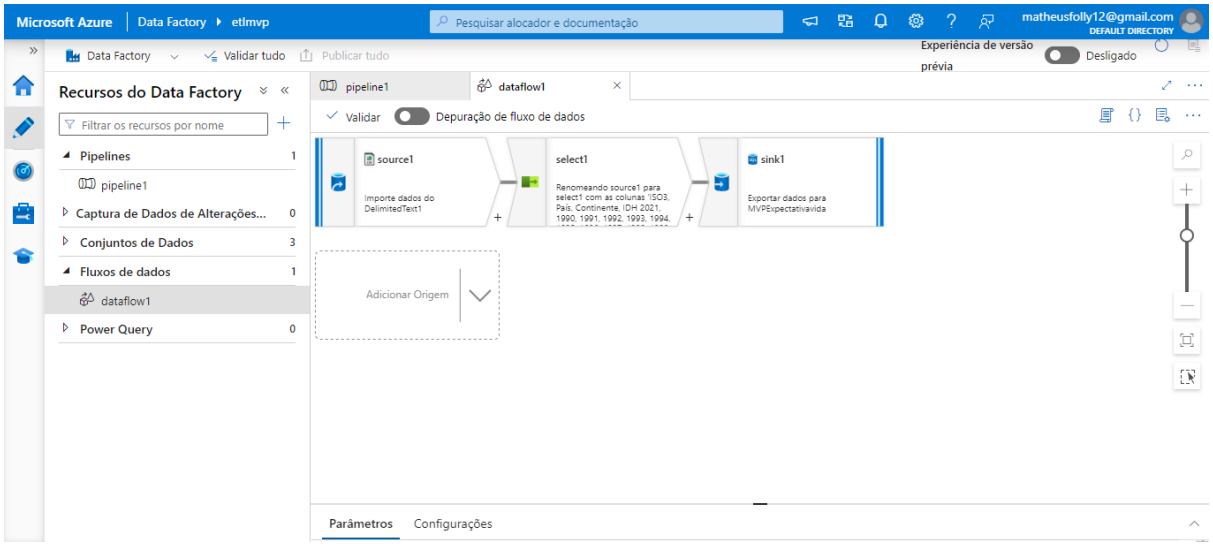


É importante ressaltar que a coluna “IDH 2021” trata-se da posição do país no ranking do IDH de 2021. Ela não traz propriamente o índice de desenvolvimento de cada país, apenas sinaliza qual foi a posição no ranking, este posiciona de 1 a 195 os países, sendo a 1ª posição o país com o maior índice de desenvolvimento humano e a posição 195 o país com menor índice de desenvolvimento humano.

As colunas de 1990 a 2021, referentes a expectativa de vida, não consideram os diferentes gêneros, são uma média de expectativa de vida para o ser humano nascido no ano citado em um determinado país.

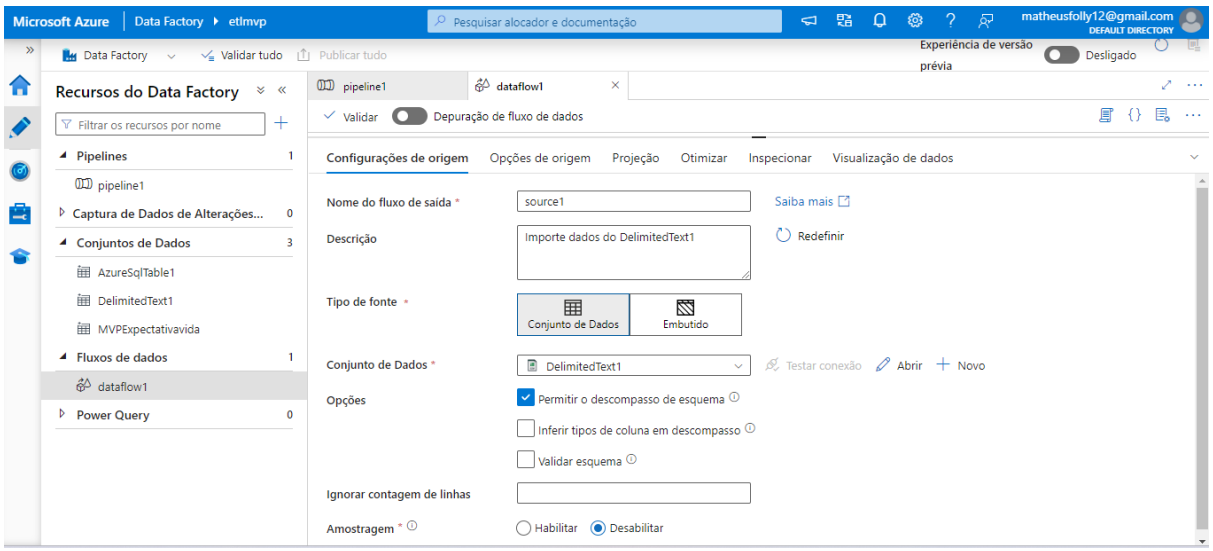
4. Carga

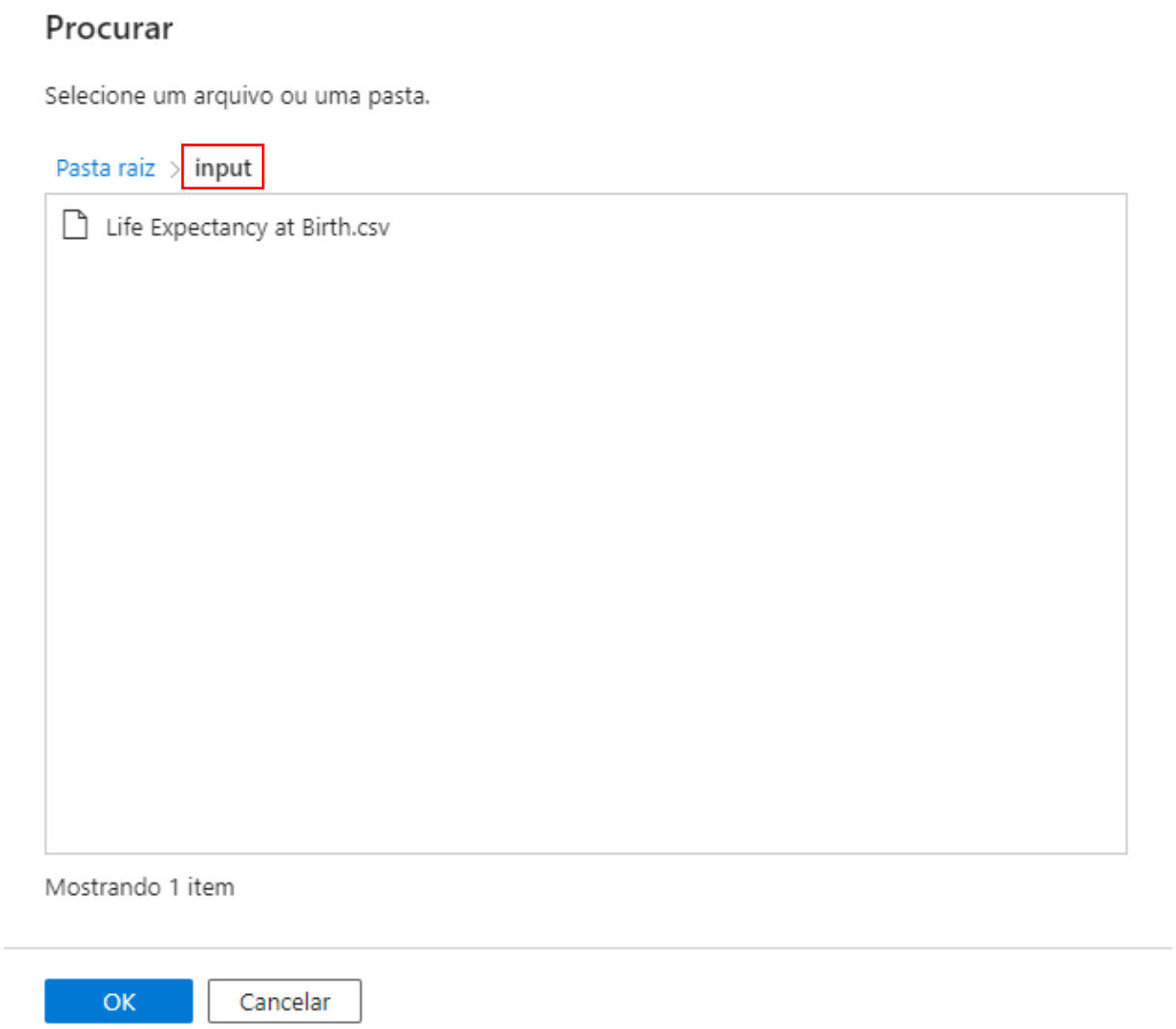
Para o ETL, criei o job nomeado “etlmvp” utilizando o serviço Data Factories:



- Extração:

As configurações foram feitas para extrair o arquivo do contêiner “input”:





- Transformação:

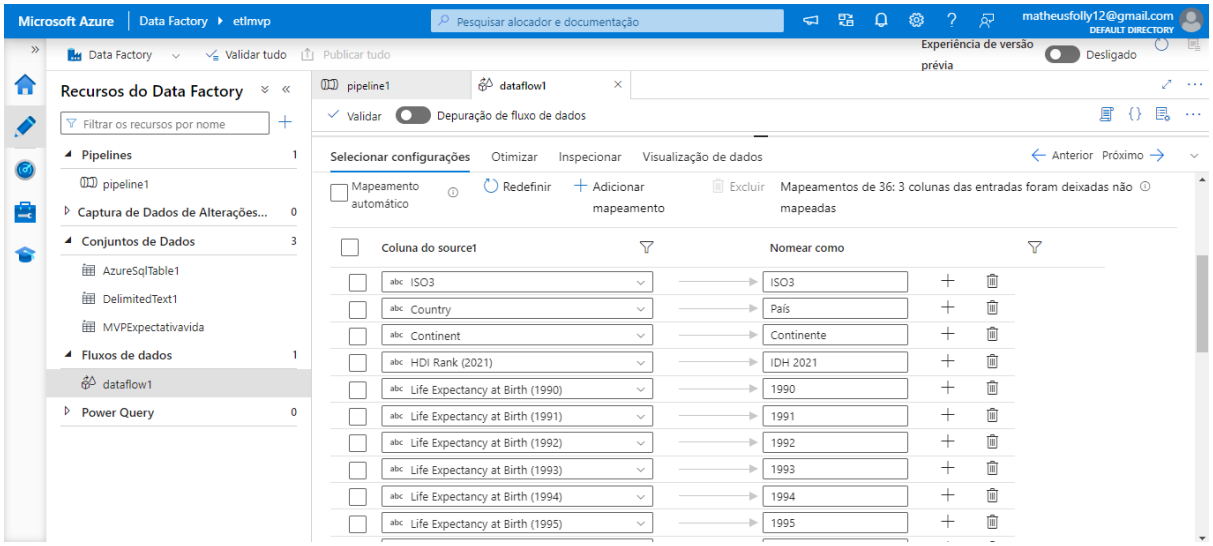
Na etapa de transformação:

- Excluí as colunas Hemisphere, Human Development Groups, UNDP Developing Regions, essas não farão parte da análise.

- Alterei o nome das colunas Country e Continent para a tradução em português.

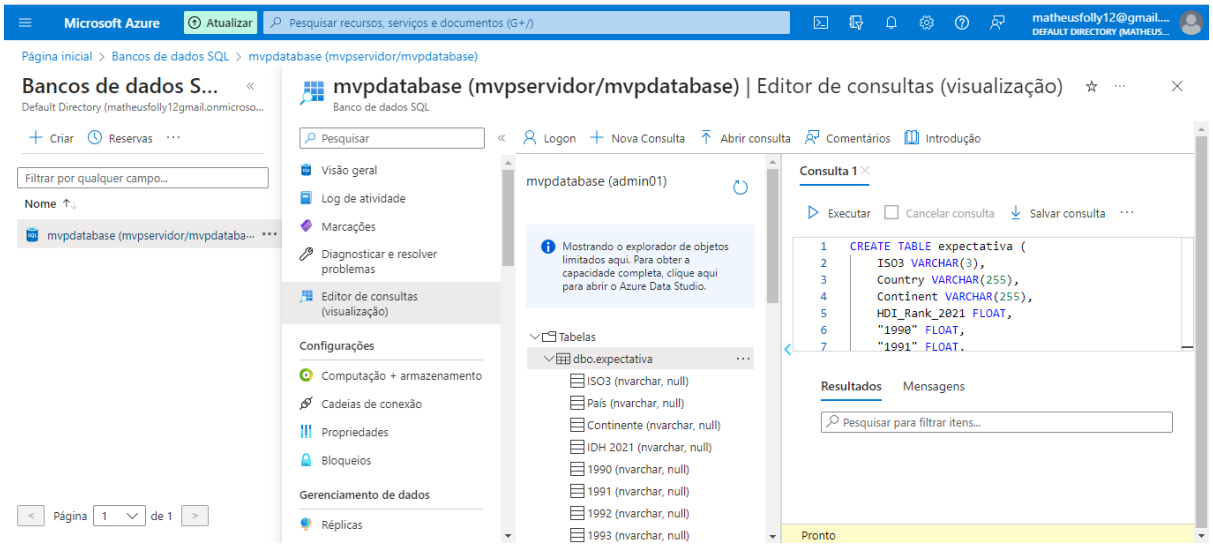
- Alterei o nome da coluna “HDI Rank (2021)” para IDH 2021, que seria a tradução da sigla para o português.

- Alterei o nome de todas as colunas que representam os dados das expectativas de vida. Inicialmente estavam como “Life Expectancy at Birth (ano x)”, renomeei com apenas o ano em questão. Acredito que desta forma facilitará a visualização e consultas futuras.

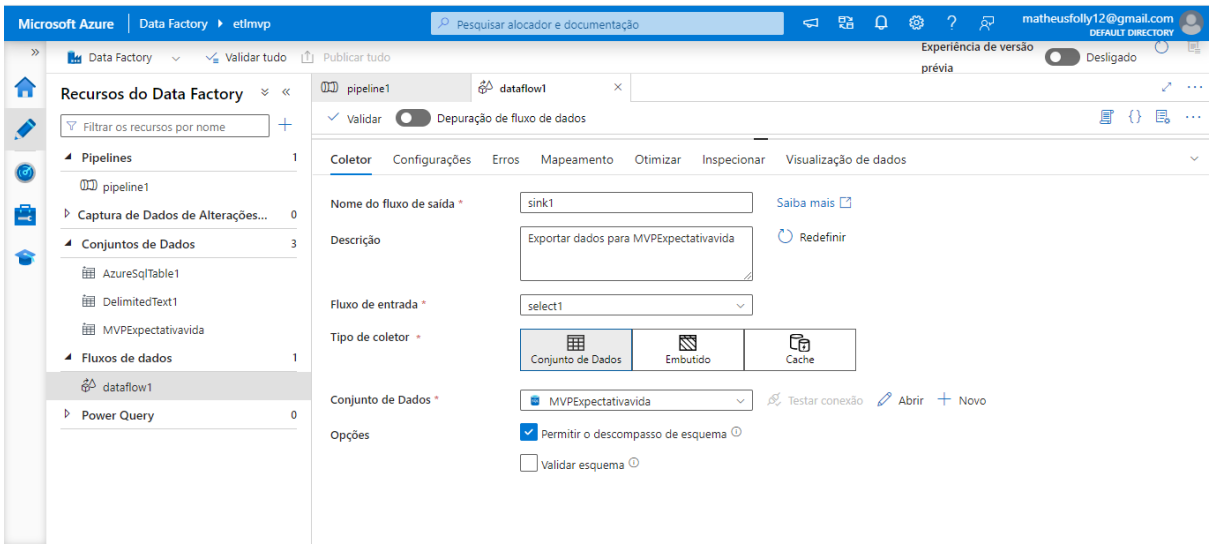


- Carga:

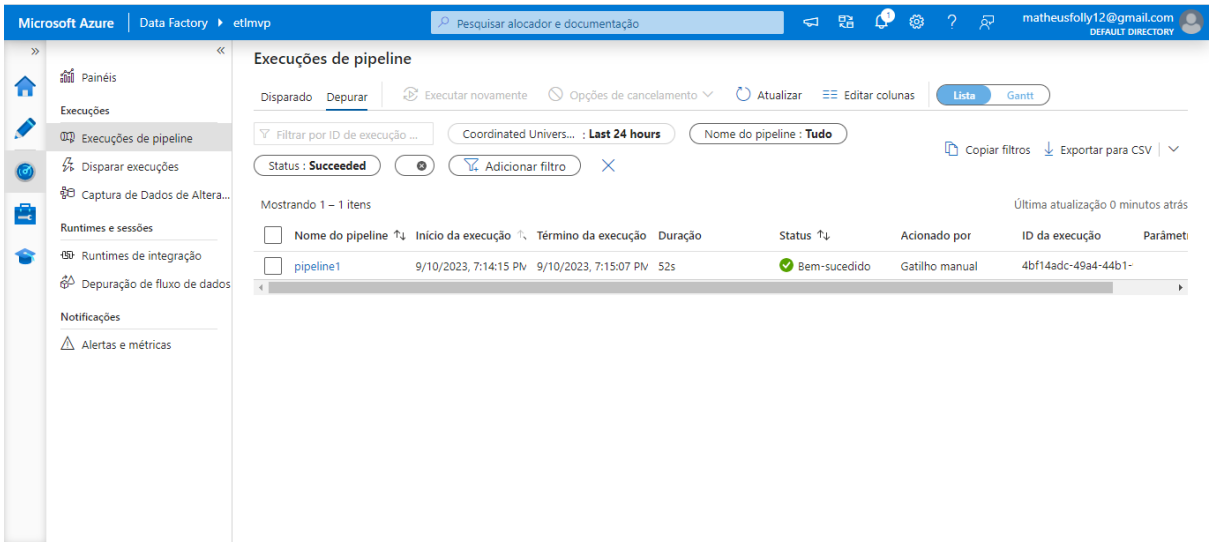
Antes da carga, criei um banco de dados chamado “mvpdatabase”, onde criei a tabela “expectativa”, por meio do código CREATE TABLE:



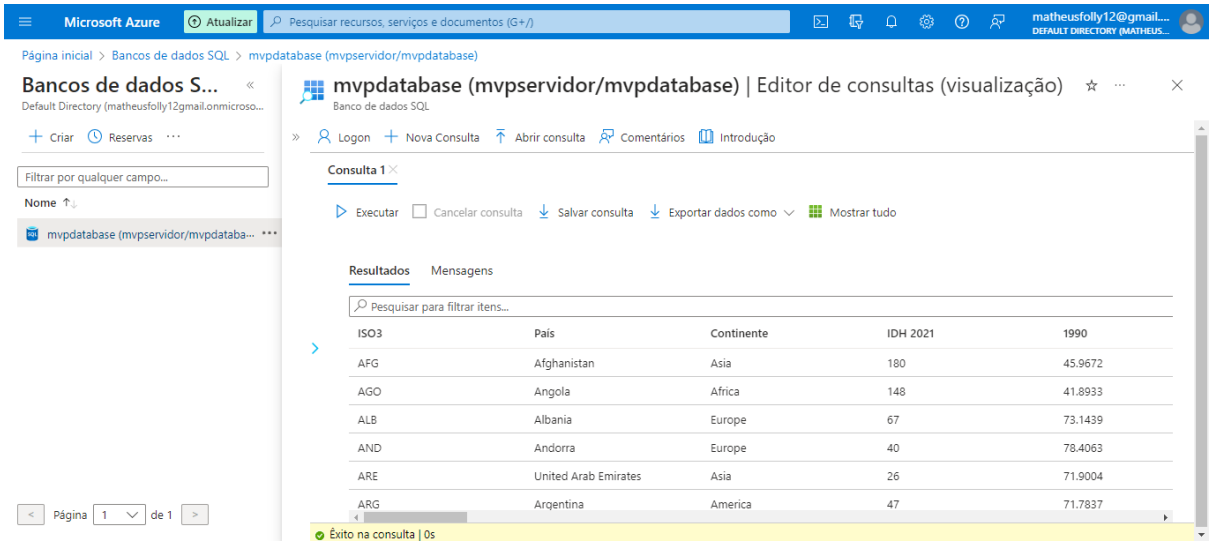
Depois disso, fiz as configurações necessárias para carregar os dados transformados para a tabela criada:



Execução bem sucedida:



Dados carregados para a tabela alvo:



A análise do trabalho foi feita por meio do Jupyter Notebook. Para conectar o Jupyter no banco de dados SQL da Azure, rodei um código utilizando o link do servidor, o nome do database, o usuário e senha. Os dados permaneceram na nuvem durante todo o trabalho.

```
conn_str = f'DRIVER={driver};SERVER={server};DATABASE={database};UID={username};PWD={password}'
```

```
conn = pyodbc.connect(conn_str)
```

```
query = 'SELECT * FROM expectativa'
```

```
df = pd.read_sql(query, conn)
```

Não incluí os dados reais do database no print por motivos de segurança.

5. Análise

- **Qualidade dos dados**

➢ Ausência de valor:

```
In [12]: linhas_em_branco = df['IDH 2021'].isna()
         linhas_com_idh_em_branco = df[linhas_em_branco]
         display(linhas_com_idh_em_branco)
```

	ISO3	País	Continente	IDH 2021	1990	1991	1992	1993	1994	1995	...	2012	2013	2014	2015	2016	2017	2018
108	MCO	Monaco	Europe	None	75.8078	76.3011	76.8424	77.4391	78.1133	78.7799	...	84.5293	84.8656	84.8804	85.2585	85.7228	85.6525	86.4643
132	NRU	Nauru	Oceania	None	61.3557	61.3031	61.2233	61.3376	61.4626	61.3214	...	60.7152	61.2314	61.8074	62.269	62.7018	62.988	63.2337
142	PRK	North Korea	Asia	None	70.21	70.3125	70.8239	71.8705	70.7181	60.8944	...	71.6261	72.3189	72.9423	72.7844	72.8045	72.9778	73.0309
158	SOM	Somalia	Africa	None	47.1055	26.5647	27.3116	50.6485	50.3266	50.6005	...	53.1611	53.8433	54.2773	54.857	55.0444	55.6536	56.3754

4 rows × 36 columns

Há 4 linhas com ausência de valor na coluna “IDH 2021”. Busquei os valores na base da UNDP (de onde o autor do database no Kaggle retirou os dados), os países realmente estão sem valores para o IDH e sem colocação no ranking. Por algum motivo os dados destes países não foram coletados. Inicialmente, estes valores nulos não afetarão a análise, tratarei os mesmos quando for necessário.

➢ Linhas Duplicadas:

```
In [13]: linhas_duplicadas = df.duplicated()
         print(linhas_duplicadas.sum())

         0
```

Não há linhas duplicadas.

➢ Tratamento de valores:

Inicialmente, foi observado que a base de dados não realiza a separação dos subcontinentes da América:

```
In [9]: display(df.loc[df['Continente'] == 'America'].head(5))
```

	ISO3	País	Continente	IDH 2021	1990	1991	1992	1993	1994	1995	...	2012	2013	2014	2015	2016	2017	2018
5	ARG	Argentina	America	47	71.7837	72.319	72.4295	72.5646	73.1725	73.1333	...	76.4669	76.4908	76.7549	76.7602	76.3077	76.833	76.9994
7	ATG	Antigua and Barbuda	America	71	73.4922	73.4354	73.4168	73.4819	73.5912	73.6363	...	77.3502	77.5834	77.8577	77.9127	78.1516	78.2683	78.5112
18	BHS	Bahamas	America	55	70.1331	69.6157	70.6679	70.6559	70.8507	70.7628	...	72.7528	73.0237	73.3658	73.104	73.5368	73.6317	73.8057
21	BLZ	Belize	America	123	70.7437	70.2484	69.9793	69.9882	70.3164	69.8874	...	73.245	73.6661	73.3107	73.1866	73.3988	73.5624	73.7031
22	BOL	Bolivia	America	118	56.4221	57.0973	57.6287	58.4179	58.9886	59.5337	...	66.7046	67.0207	67.1632	67.3182	67.6277	67.7015	67.748

5 rows × 36 columns

Para refinar a análise, separei a América em Norte, Central e Sul:

```
In [10]: df2 = df.copy()

regras_continente = {
'Canada': 'North America',
'United States': 'North America',
'Mexico': 'North America',
'Antigua and Barbuda': 'Central America',
'Bahamas': 'Central America',
'Barbados': 'Central America',
'Belize': 'Central America',
'Costa Rica': 'Central America',
'Cuba': 'Central America',
'Dominica': 'Central America',
'El Salvador': 'Central America',
'Grenada': 'Central America',
'Guatemala': 'Central America',
'Haiti': 'Central America',
'Honduras': 'Central America',
'Jamaica': 'Central America',
'Nicaragua': 'Central America',
'Panama': 'Central America',
'Dominican Republic': 'Central America',
'Saint Lucia': 'Central America',
'Saint Kitts and Nevis': 'Central America',
'Saint Vincent and the Grenadines': 'Central America',
'Trinidad and Tobago': 'Central America',
'Argentina': 'South America',
'Bolivia': 'South America',
'Brazil': 'South America',
'Chile': 'South America',
```

```
'Colombia': 'South America',
'Ecuador': 'South America',
'Guyana': 'South America',
'Paraguay': 'South America',
'Peru': 'South America',
'Suriname': 'South America',
'Uruguay': 'South America',
'Venezuela': 'South America'
}

for país, novo_continente in regras_continente.items():
    df2.loc[df2['País'] == país, 'Continente'] = novo_continente


nomes_paises = ['Brazil', 'Bahamas','United States','Chile']
display(df2[df2['País'].isin(nomes_paises)])
```

	ISO3	País	Continente	IDH 2021	1990	1991	1992	1993	1994	1995	...	2012	2013	2014	2015	2016	2017	2018
18	BHS	Bahamas	Central America	55	70.1331	69.6157	70.6679	70.6559	70.8507	70.7628	...	72.7528	73.0237	73.3658	73.104	73.5368	73.6317	73.8057
23	BRA	Brazil	South America	87	65.9848	66.3096	66.7082	67.1092	67.5684	67.9191	...	73.5517	73.9185	74.3058	74.3325	74.4415	74.8266	75.1095
31	CHL	Chile	South America	42	72.574	73.5741	74.1354	74.1953	74.6139	74.6085	...	79.0233	79.3391	79.4726	79.746	80.0794	80.3501	80.1334
184	USA	United States	North America	21	75.3699	75.5227	75.7776	75.567	75.7377	75.8536	...	78.9441	78.9507	79.0175	78.8694	78.8482	78.8213	78.9896

4 rows × 36 columns

Além deste tratamento, as colunas numéricas inicialmente consideradas como tipo object, foram convertidas para Float:

```
In [11]: colunas_converter = ['IDH 2021', '1990', '1991', '1992', '1993', '1994', '1995', '1996', '1997', '1998',
                              '1999', '2000', '2001', '2002', '2003', '2004', '2005', '2006', '2007', '2008', '2009',
                              '2010', '2011', '2012', '2013', '2014', '2015', '2016', '2017', '2018', '2019', '2020', '2021']

         for coluna in colunas_converter:
             df2[coluna] = df2[coluna].astype(float)

         print(df2.dtypes)
```

```
ISO3          object
País          object
Continente    object
IDH 2021     float64
1990         float64
1991         float64
1992         float64
1993         float64
1994         float64
1995         float64
1996         float64
1997         float64
1998         float64
1999         float64
2000         float64
2001         float64
2002         float64
2003         float64
2004         float64
2005         float64
```

De maneira geral, a base de dados analisada apresentou poucos problemas em relação a qualidade dos dados.

- **Solução do problema**

➢ Qual a média de expectativa de vida por continente em 2021?

```
In [21]: media_continente2021 = df2.groupby('Continente')['2021'].mean()

         print("Média de Expectativa de Vida por continente em 2021:")
         display(media_continente2021)
```

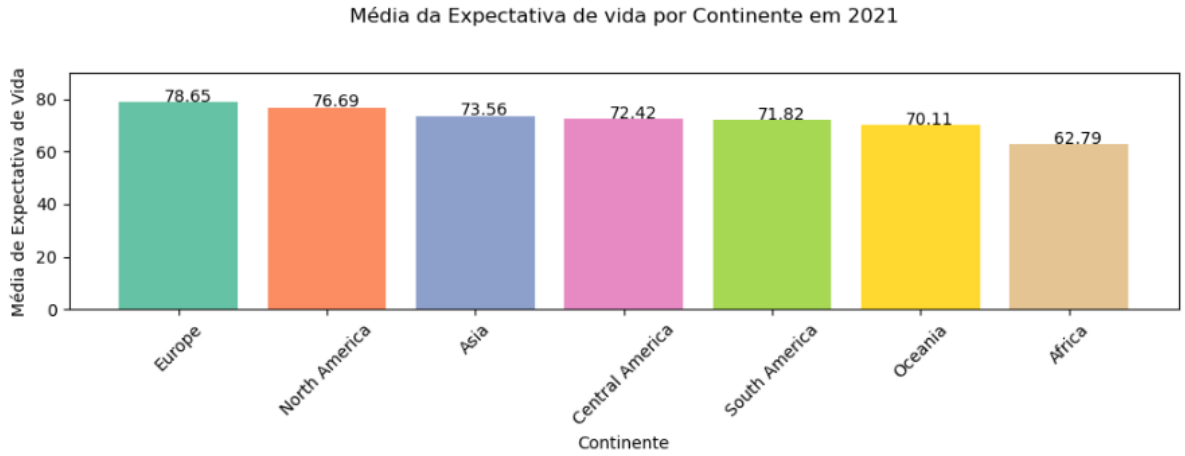
```
Média de Expectativa de Vida por continente em 2021:

Continente
Africa             62.785587
Asia               73.563327
Central America    72.423055
Europe             78.650456
North America      76.689133
Oceania            70.112307
South America      71.815775
Name: 2021, dtype: float64
```

Verifiquei se há algum outlier distorcendo a média por continente. Multipliquei por 1.5 o intervalo quartil para localizar valores que ultrapassam o tamanho de 50% deste intervalo, para cima ou para baixo. Não foram encontrados valores com essas especificidades.

```
In [25]: #Localizar Outliers

         Q1 = df2['2021'].quantile(0.25)
         Q3 = df2['2021'].quantile(0.75)

         IQR = Q3 - Q1

         limite_inferior = Q1 - 1.5 * IQR
         limite_superior = Q3 + 1.5 * IQR

         outliers = df2[(df2['2021'] < limite_inferior) | (df2['2021'] > limite_superior)]

         print(outliers[['País', 'Continente', '2021']])
```

```
Empty DataFrame
Columns: [País, Continente, 2021]
Index: []
```

```
In [16]: media_continente2021 = media_continente2021.sort_values(ascending=False)

         plt.figure(figsize=(10, 4))
         cores = sns.color_palette("Set2", len(media_continente2021))

         bars = plt.bar(media_continente2021.index, media_continente2021.values, color=cores)

         for bar, valor in zip(bars, media_continente2021.values):
             plt.text(bar.get_x() + bar.get_width() / 2 - 0.1, bar.get_height() + 0.2, f'{valor:.2f}', fontsize=10)

         plt.title('Média da Expectativa de vida por Continente em 2021', pad=30)
         plt.xlabel('Continente')
         plt.ylabel('Média de Expectativa de Vida')

         plt.ylim(0, 90)


         plt.xticks(rotation=45)
         plt.tight_layout()
         plt.show()
```



Pelo gráfico podemos observar que a Europa tem a maior média de expectativa de vida. Observamos também, uma grande diferença entre a maior e a menor média. Em 2021, uma pessoa nascida na Europa, vive em média 15 anos a mais que uma pessoa nascida na África.

- Quais são os 10 países com melhores posições no IDH 2021? A quais continentes pertencem?

```
In [18]: df_corr = df2[['País','IDH 2021','Continente','2021']]
         df_corr = df_corr.sort_values(by='IDH 2021', ascending=True)

         df_corr.rename(columns={"IDH 2021": "Ranking IDH 2021","2021": "Expectativa de Vida em 2021"}, inplace=True)

         display(df_corr.loc[df_corr['Ranking IDH 2021']<11])
```

	País	Ranking IDH 2021	Continente	Expectativa de Vida em 2021
30	Switzerland	1.0	Europe	83.9872
130	Norway	2.0	Europe	83.2339
82	Iceland	3.0	Europe	82.6782
72	Hong Kong	4.0	Asia	85.4734
8	Australia	5.0	Oceania	84.5265
47	Denmark	6.0	Europe	81.3753
165	Sweden	7.0	Europe	82.9833
79	Ireland	8.0	Europe	81.9976
44	Germany	9.0	Europe	80.6301
129	Netherlands	10.0	Europe	81.6873

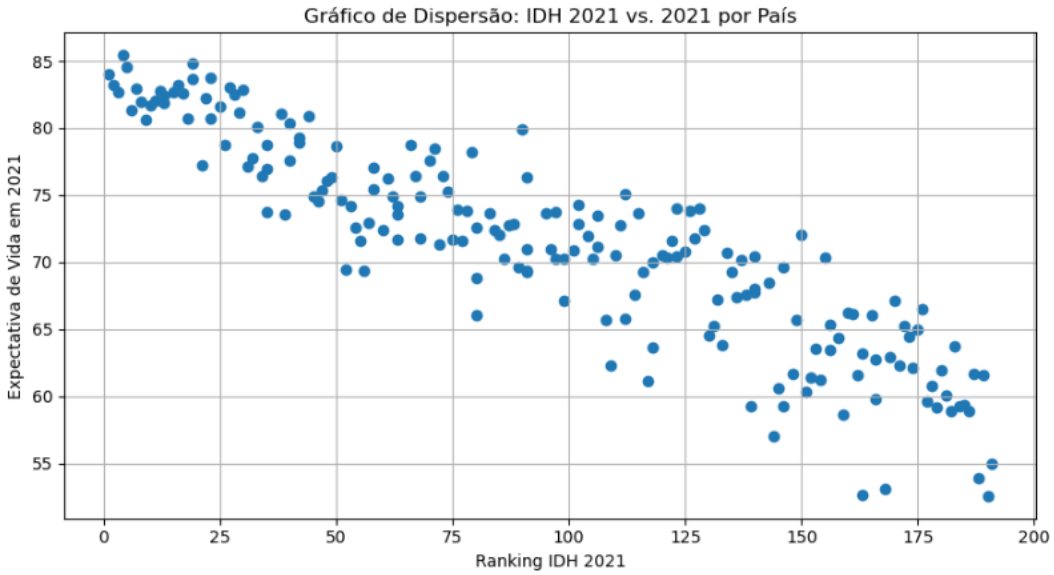
Os 10 países mais bem posicionados no ranking do IDH 2021 são respectivamente: Suíça, Noruega, Islândia, Hong Kong, Austrália, Dinamarca, Suécia, Irlanda, Alemanha e Países Baixos. Dos 10 países citados 8 estão na Europa.

Desta forma é possível observar a predominância de países Europeus entre os países com melhores índices de desenvolvimento humano. Como foi visto anteriormente, a Europa também tem a maior média de expectativa de vida em relação aos outros continentes. Há alguma correlação evidente entre o melhor IDH e a maior expectativa de vida? Este ponto será analisado a seguir.

- Qual a correlação entre a expectativa de vida e o ranking do IDH em 2021?

Para analisar esta correlação, irei desconsiderar os 4 países com IDH de valor nulo vistos anteriormente.

```
In [18]: df_corr = df_corr.dropna(subset=['Ranking IDH 2021']) # excluí os valores nulos

correlacao = df_corr['Expectativa de Vida em 2021'].corr(df_corr['Ranking IDH 2021'])

idh_2021 = df_corr["Ranking IDH 2021"]
ano_2021 = df_corr["Expectativa de Vida em 2021"]


plt.figure(figsize=(9, 5))
plt.scatter(idh_2021, ano_2021, marker='o')


plt.xlabel("Ranking IDH 2021")
plt.ylabel("Expectativa de Vida em 2021")
plt.title("Gráfico de Dispersão: IDH 2021 vs. 2021 por País")


plt.grid(True)
plt.tight_layout()
plt.show()

print(f'A correlação entre a expectativa de vida e o ranking do IDH é: {round(correlacao,2)}')
```




```
A correlação entre a expectativa de vida e o ranking do IDH é: -0.9
```

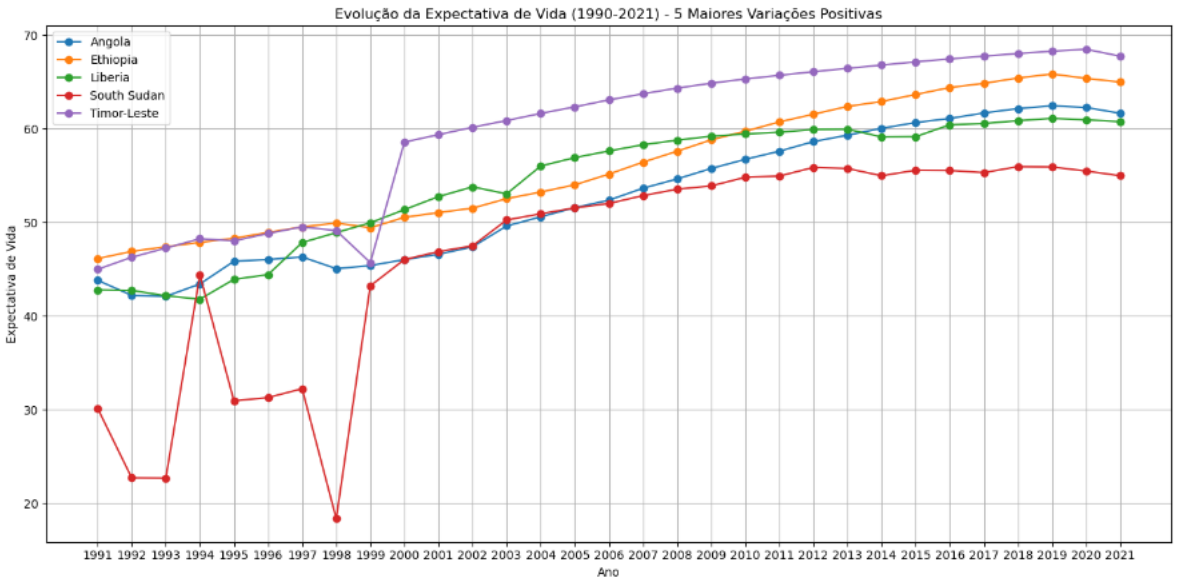
Conforme observado, a correlação encontrada entre a Expectativa de Vida de 2021 e o Ranking do IDH de 2021 é -0,90. Este número indica uma forte correlação negativa. A medida que a expectativa de vida aumenta, a posição no ranking do IDH diminui, ou seja, mais bem colocado é o país no ranking. Lembrando que o ranking do IDH posiciona os países de 1 a 195, sendo a 1ª posição o país com o maior índice de desenvolvimento humano e a posição 195 o país com menor índice de desenvolvimento humano.

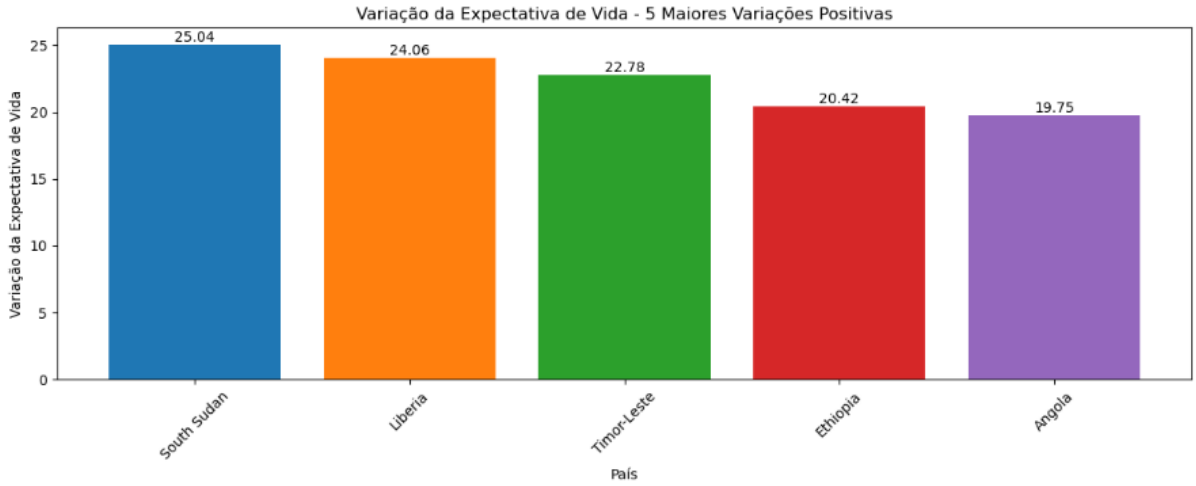
Esta análise indica que em 2021, uma expectativa de vida alta pode estar relacionada a um IDH elevado. O gráfico de dispersão ilustra bem a variação de idades e sua comparação com as posições no ranking do IDH de 2021.

- Quais países tiveram as maiores variações positivas na expectativa de vida entre 1990 a 2021? Algum desses países está entre os 10 primeiros colocados no ranking do IDH 2021?

In [19]:
```
df2['Variação_Expectativa_de_Vida'] = df2['2021'] - df2['1990']

Variação = 'Variação_Expectativa_de_Vida'
maiores_var = df2.nlargest(5, Variação)

# Gráfico Linhas

paises_maiores_var = maiores_var['País'].tolist()

df_paises_maiores_var = df2[df2['País'].isin(paises_maiores_var)]

colunas_anos = df_paises_maiores_var.columns[5:-1]

df_paises_maiores_var = df_paises_maiores_var.set_index('País')[colunas_anos].T

plt.figure(figsize=(14, 7))
for pais in df_paises_maiores_var.columns:
    plt.plot(df_paises_maiores_var.index, df_paises_maiores_var[pais], marker='o', linestyle='-', label=pais)

plt.title('Evolução da Expectativa de Vida (1990-2021) - 5 Maiores Variações Positivas')
plt.xlabel('Ano')
plt.ylabel('Expectativa de Vida')
plt.grid(True)
plt.legend(loc='upper left')
plt.tight_layout()

plt.show()
```

```
# Gráfico Colunas
plt.figure(figsize=(12, 5))
for i, pais in enumerate(paises_maiores_var):
    variacao = df2.loc[df2['País'] == pais, 'Variação_Expectativa_de_Vida'].values[0]
    plt.bar(pais, variacao)

plt.title('Variação da Expectativa de Vida - 5 Maiores Variações Positivas')
plt.xlabel('País')
plt.ylabel('Variação da Expectativa de Vida')
plt.xticks(rotation=45)
plt.tight_layout()

# Rótulo de dados
for i, pais in enumerate(paises_maiores_var):
    variacao = df2.loc[df2['País'] == pais, 'Variação_Expectativa_de_Vida'].values[0]
    plt.text(i, variacao, f'{variacao:.2f}', ha='center', va='bottom')

plt.show()
```






```
# Países com as maiores variações positivas na expectativa de vida

paises_maioresvariacoes = df2.loc[df2['País'].isin(['South Sudan', 'Liberia','Timor-Leste','Ethiopia','Angola'])]
paises_maioresvariacoes = paises_maioresvariacoes.sort_values(by='2021', ascending=False)

display(paises_maioresvariacoes[['País','Continente','IDH 2021','2021']])
```

	País	Continente	IDH 2021	2021
174	Timor-Leste	Asia	140.0	67.7369
55	Ethiopia	Africa	175.0	64.9748
1	Angola	Africa	148.0	61.6434
98	Liberia	Africa	178.0	60.7472
160	South Sudan	Africa	191.0	54.9752

Os gráficos acima mostram os países que mais aumentaram sua expectativa de vida entre 1990 a 2021. Isto não significa que esses países atingiram uma expectativa de idade elevada. A alta variação ocorreu devido a uma expectativa de vida abaixo do comum por volta de 1990 em tais países. Esta baixa longevidade pode estar relacionada a conflitos bélicos que ocorriam na época:

Sudão do Sul: 1983 a 2005 - Segunda Guerra Civil Sudanesa,

Libéria: 1989 a 1996 - Primeira Guerra civil da Libéria,

Timor Leste: 1975 a 1999 - Luta pela Independência,

Etiópia: 1974 a 1991 - Guerra Civil da Etiópia,

Angola: 1975 a 2002 - Guerra Civil Angolana.

Mesmo com uma variação positiva de 19 a 25 anos na expectativa de vida ao longo do tempo, os países citados ainda possuem uma expectativa de idade baixa. Estes países estão com posições superiores a 140 no ranking de IDH 2021.

6. Conclusão

A partir da análise feita, é possível concluir que tanto as maiores expectativas de vida quanto os melhores índices de desenvolvimento humano pertencem, predominantemente, a países europeus. Tal afirmação não se trata de coincidência. Conforme foi analisado, existe uma correlação entre as maiores expectativas de vida e as melhores posições no ranking de IDH.

Esta correlação também não é originada do acaso. O cálculo do índice de desenvolvimento humano é feito a partir de 3 componentes: PIB, educação e expectativa de vida. A própria expectativa de vida é considerada no cálculo do IDH. O índice de desenvolvimento humano foi criado com a proposta de ser uma medida geral de desenvolvimento, contrapondo o PIB que considera apenas a dimensão econômica.

Foi possível verificar também as diferenças na longevidade de pessoas nascidas em diferentes continentes. Com uma diferença de 15 anos entre a menor e maior média de expectativa de vida. Tal diferença pode estar relacionada à condições sanitárias e mortes precoces, fatores que interferem diretamente na expectativa de vida.

Por fim, foi analisado quais países tiveram o maior aumento na expectativa de vida entre os anos de 1990 a 2021 e se esses países estavam nas melhores posições do ranking de IDH 2021. O objetivo era verificar se algum dos países entre os melhores IDHs já tiveram uma baixa expectativa de vida e como foi a evolução. A partir da análise, verifiquei que os países que mais variaram positivamente não possuem uma alta expectativa de vida e nem um bom IDH. As variações positivas desses países se originaram de uma expectativa de vida abaixo do normal por volta de 1990, provavelmente relacionadas a mortes precoces em guerras que aconteciam na época. Com o tempo a longevidade desses países aumentou, uma grande variação entre 19 a 25 anos, mas continua abaixo da média, quando comparamos com outros países.

7. Autoavaliação

O objetivo proposto foi concluído, consegui responder aos problemas por meio da análise feita. A maior parte das dificuldades encontradas foram em relação aos serviços de nuvem. Inicialmente, havia começado o trabalho utilizando a AWS, porém não consegui estabelecer a conexão entre o AWS Glue e o Redshift para a carga dos dados tratados. Aparecia um erro genérico dizendo que a conexão havia falhado. Para contornar o problema, resolvi realizar o trabalho pela Microsoft Azure, no começo houve a dificuldade de ser a primeira vez utilizando a plataforma, mas com testes e pesquisas, consegui concluir o ETL e carregar os dados para um banco de dados SQL da própria Azure.

Ainda não passei pelas Sprints de Análise de Dados e Machine Learning, mas resolvi realizar a análise do MVP utilizando Python como um desafio. Já havia um conhecimento prévio e acredito que aprendo melhor na prática. Tive alguns erros com códigos no começo, mas com insistência e pesquisas, consegui alcançar o resultado esperado.

Para um futuro trabalho, acredito que seria interessante analisar de forma mais específica países escolhidos, trazendo outros dados como: PIB, educação e nível de desemprego. De forma a traçar a trajetória do país ao longo do tempo, mostrando sua evolução ou não.

8. Referências Bibliográficas

https://www.kaggle.com/datasets/iamsouravbanerjee/life-expectancy-at-birth-across-the-globe

https://hdr.undp.org/system/files/documents/global-report-document/hdr2021-22ptpdf.pdf

https://www.undp.org/pt/brazil/o-que-%C3%A9-o-idh

https://educacao.uol.com.br/disciplinas/geografia/indice-de-desenvolvimento-humano-idh-mede-nivel-de-qualidade-de-vida.htm#:~:text=O%20c%C3%A1lculo%20do%20IDH%20leva,todos%20os%20n%C3%ADveis%20de%20ensino.

https://www.cnnbrasil.com.br/internacional/

https://mundoeducacao.uol.com.br/geografia/os-conflitos-na-africa.htm

https://brasilescola.uol.com.br/geografia/conflitos-na-africa.htm

https://www.infopedia.pt/apoio/artigos/$luta-pela-independencia-de-timor-leste

https://www.publico.pt/2021/06/30/p3/fotogaleria/liberia-chora-silencio-ensurdecedor-trauma-guerra-406364