Etapa 1 - Importação das bibliotecas necessárias Importamos as bibliotecas pandas para manipulação de dados e os para listar os arquivos.

```
import pandas as pd
import os
```

Etapa 2 - Upload dos arquivos para o Google Colab Solicitamos o upload dos 17 arquivos .xlsx, seguindo o padrão de nomenclatura:

ICJ_2008.xlsx ICJ_2009.xlsx ... ICJ_2024.xlsx

```
from google.colab import files
uploaded = files.upload()
```

Escolher arquivos 17 arquivos

- **ICJ_2008.xlsx**(application/vnd.openxmlformats-officedocument.spreadsheetml.sheet) - 10859178 bytes, last modified: 24/03/2025 - 100% done
- **ICJ_2009.xlsx**(application/vnd.openxmlformats-officedocument.spreadsheetml.sheet) - 10859178 bytes, last modified: 24/03/2025 - 100% done
- **ICJ_2010.xlsx**(application/vnd.openxmlformats-officedocument.spreadsheetml.sheet) - 10859178 bytes, last modified: 24/03/2025 - 100% done
- **ICJ_2011.xlsx**(application/vnd.openxmlformats-officedocument.spreadsheetml.sheet) - 10859178 bytes, last modified: 24/03/2025 - 100% done
- **ICJ_2012.xlsx**(application/vnd.openxmlformats-officedocument.spreadsheetml.sheet) - 10859178 bytes, last modified: 24/03/2025 - 100% done
- **ICJ_2013.xlsx**(application/vnd.openxmlformats-officedocument.spreadsheetml.sheet) - 10859178 bytes, last modified: 24/03/2025 - 100% done
- **ICJ_2014.xlsx**(application/vnd.openxmlformats-officedocument.spreadsheetml.sheet) - 10859178 bytes, last modified: 24/03/2025 - 100% done
- **ICJ_2015.xlsx**(application/vnd.openxmlformats-officedocument.spreadsheetml.sheet) - 10859178 bytes, last modified: 24/03/2025 - 100% done
- **ICJ_2016.xlsx**(application/vnd.openxmlformats-officedocument.spreadsheetml.sheet) - 10859178 bytes, last modified: 24/03/2025 - 100% done
- **ICJ_2017.xlsx**(application/vnd.openxmlformats-officedocument.spreadsheetml.sheet) - 10859178 bytes, last modified: 24/03/2025 - 100% done
- **ICJ_2018.xlsx**(application/vnd.openxmlformats-officedocument.spreadsheetml.sheet) - 10859178 bytes, last modified: 24/03/2025 - 100% done
- **ICJ_2019.xlsx**(application/vnd.openxmlformats-officedocument.spreadsheetml.sheet) - 10859178 bytes, last modified: 24/03/2025 - 100% done
- **ICJ_2020.xlsx**(application/vnd.openxmlformats-officedocument.spreadsheetml.sheet) - 10859178 bytes, last modified: 24/03/2025 - 100% done
- **ICJ_2021.xlsx**(application/vnd.openxmlformats-officedocument.spreadsheetml.sheet) - 10859178 bytes, last modified: 24/03/2025 - 100% done
- **ICJ_2022.xlsx**(application/vnd.openxmlformats-officedocument.spreadsheetml.sheet) - 10859178 bytes, last modified: 24/03/2025 - 100% done
- **ICJ_2023.xlsx**(application/vnd.openxmlformats-officedocument.spreadsheetml.sheet) - 10859178 bytes, last modified: 24/03/2025 - 100% done
- **ICJ_2024.xlsx**(application/vnd.openxmlformats-officedocument.spreadsheetml.sheet) - 10859178 bytes, last modified: 24/03/2025 - 100% done

```
Saving ICJ_2008.xlsx to ICJ_2008 (1).xlsx
Saving ICJ_2009.xlsx to ICJ_2009 (1).xlsx
Saving ICJ_2010.xlsx to ICJ_2010 (1).xlsx
Saving ICJ_2011.xlsx to ICJ_2011 (1).xlsx
Saving ICJ_2012.xlsx to ICJ_2012 (1).xlsx
Saving ICJ_2013.xlsx to ICJ_2013 (1).xlsx
Saving ICJ_2014.xlsx to ICJ_2014 (1).xlsx
Saving ICJ_2015.xlsx to ICJ_2015 (1).xlsx
Saving ICJ_2016.xlsx to ICJ_2016 (1).xlsx
Saving ICJ_2017.xlsx to ICJ_2017 (1).xlsx
Saving ICJ_2018.xlsx to ICJ_2018 (1).xlsx
Saving ICJ_2019.xlsx to ICJ_2019 (1).xlsx
Saving ICJ_2020.xlsx to ICJ_2020 (1).xlsx
Saving ICJ_2021.xlsx to ICJ_2021 (1).xlsx
Saving ICJ_2022.xlsx to ICJ_2022 (1).xlsx
Saving ICJ_2023.xlsx to ICJ_2023 (1).xlsx
Saving ICJ_2024.xlsx to ICJ_2024 (1).xlsx
```

Etapa 3 - Carregamento e tratamento dos arquivos Criamos uma função que:

Lê o arquivo Excel. Adiciona uma coluna com o respectivo ano (extraído do nome do arquivo). Retorna o DataFrame tratado. Iteramos sobre os anos de 2008 a 2024, carregando e empilhando os dados.

```
def carregar_arquivo_com_ano(nome_arquivo, ano):
    # Lê a planilha, pulando a primeira linha (vazia), e usando a segunda como cabeçalho
    df = pd.read_excel(nome_arquivo, header=1)
    df['Ano'] = ano
    return df

# Lista de anos
anos = list(range(2008, 2025))

# Lista para armazenar os DataFrames
dataframes = []

# Loop para carregar todos os arquivos
for ano in anos:
    nome_arquivo = f"ICJ_{ano}.xlsx"
    print(f"Lendo o arquivo: {nome_arquivo}")
    df_ano = carregar_arquivo_com_ano(nome_arquivo, ano)
    dataframes.append(df_ano)

# Concatenar todos os DataFrames em um único banco de dados
df_total = pd.concat(dataframes, ignore_index=True)
```

```
Lendo o arquivo: ICJ_2008.xlsx
Lendo o arquivo: ICJ_2009.xlsx
Lendo o arquivo: ICJ_2010.xlsx
Lendo o arquivo: ICJ_2011.xlsx
Lendo o arquivo: ICJ_2012.xlsx
Lendo o arquivo: ICJ_2013.xlsx
Lendo o arquivo: ICJ_2014.xlsx
Lendo o arquivo: ICJ_2015.xlsx
Lendo o arquivo: ICJ_2016.xlsx
Lendo o arquivo: ICJ_2017.xlsx
Lendo o arquivo: ICJ_2018.xlsx
Lendo o arquivo: ICJ_2019.xlsx
Lendo o arquivo: ICJ_2020.xlsx
```

```
Lendo o arquivo: ICJ_2021.xlsx
Lendo o arquivo: ICJ_2022.xlsx
Lendo o arquivo: ICJ_2023.xlsx
Lendo o arquivo: ICJ_2024.xlsx
```

Etapa 4 - Verificações e validação do banco consolidado Visualizamos as primeiras linhas e informações do banco consolidado:

```
# Primeiras linhas do banco consolidado
df_total.head()
```

	#	01_Instituição	02_Programa	03_Modalidade	04_Cod Categoria Nível	05 _Área	06_Grande Área	07_Linha de Fomento	08_Sexo	09_Cor ou Raça	...	15
0	1.0	4C INNOVATION CONSULTING LTDA - ME	PROGRAMA RHAE - INOVACAO	Fixação de Recursos Humanos	G	Engenharia de Materiais e Metalúrgica	Engenharias	APOIO A PROJETOS DE PESQUISA	Masculino	Amarela	...	
1	2.0	A.G.L. - Incorporadora Ltda	PROGRAMA RHAE - INOVACAO	Especialista Visitante	1	Engenharia Civil	Engenharias	APOIO A PROJETOS DE PESQUISA	Masculino	Branca	...	
2	3.0	A.G.L. - Incorporadora Ltda	PROGRAMA RHAE - INOVACAO	Fixação de Recursos Humanos	F	Engenharia Civil	Engenharias	APOIO A PROJETOS DE PESQUISA	Masculino	Branca	...	
3	4.0	A.G.L. - Incorporadora Ltda	PROGRAMA RHAE - INOVACAO	Fixação de Recursos Humanos	G	Engenharia Civil	Engenharias	APOIO A PROJETOS DE PESQUISA	Masculino	Branca	...	
4	5.0	A.G.L. - Incorporadora Ltda	PROGRAMA RHAE - INOVACAO	Fixação de Recursos Humanos	H	Engenharia Civil	Engenharias	APOIO A PROJETOS DE PESQUISA	Masculino	Branca	...	

5 rows × 25 columns

```
# Estrutura do banco de dados
df_total.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1571531 entries, 0 to 1571530
Data columns (total 25 columns):
 #   Column                 Non-Null Count    Dtype
---  ------                 --------------    -----
```

```
 0   #                      1571531 non-null  float64
 1   01_Instituição         1571531 non-null  object
 2   02_Programa            1571531 non-null  object
 3   03_Modalidade          1571531 non-null  object
 4   04_Cod Categoria Nível 459340 non-null   object
 5   05 _Área               1571531 non-null  object
 6   06_Grande Área         1571531 non-null  object
 7   07_Linha de Fomento    1571531 non-null  object
 8   08_Sexo                1571531 non-null  object
 9   09_Cor ou Raça         1571531 non-null  object
 10  10_Origem do Recurso   1571531 non-null  object
 11  11_País                1571531 non-null  object
 12  12_Região              1571531 non-null  object
 13  13_Unidade Federação   1571531 non-null  object
 14  14_Sigla UF            1571531 non-null  object
 15  15_Cidade              1571531 non-null  object
 16  16_Cod Município IBGE  1553664 non-null  float64
```