

✓


UNIVERSIDAD TÉCNICA PARTICULAR DE LOJA

PROYECTO FINAL - HERRAMIENTAS PARA INTELIGENCIA ARTIFICIAL

INTEGRANTES

- Henry Guarnizo
- Eduardo Araujo
- Astrid Medina

```
pip install pygwalker
```



Collecting pygwalker

Downloading pygwalker-0.4.8.6-py3-none-any.whl (2.2 MB)

2.2/2.2 MB

9.4 MB/s

eta 0:00:00

Collecting appdirs (from pygwalker)

Downloading appdirs-1.4.4-py2.py3-none-any.whl (9.6 kB)

Collecting arrow (from pygwalker)

Downloading arrow-1.3.0-py3-none-any.whl (66 kB)

66.4/66.4 kB

9.1 MB/s

eta 0:00:00

Collecting astor (from pygwalker)

Downloading astor-0.8.1-py2.py3-none-any.whl (27 kB)

Requirement already satisfied: cachetools in /usr/local/lib/python3.10/dist-packages (from pygwalker) (5.3.3)

Requirement already satisfied: duckdb<0.11.0,>=0.10.1 in /usr/local/lib/python3.10/dist-packages (from pygwalker) (0.10.2)

Collecting gw-dsl-parser==0.1.48a6 (from pygwalker)

Downloading gw_dsl_parser-0.1.48a6-py3-none-any.whl (952 kB)

953.0/953.0 kB

15.0 MB/s

eta 0:00:00

Requirement already satisfied: ipython in /usr/local/lib/python3.10/dist-packages (from pygwalker) (7.34.0)

Requirement already satisfied: ipywidgets in /usr/local/lib/python3.10/dist-packages (from pygwalker) (7.7.1)

Requirement already satisfied: jinja2 in /usr/local/lib/python3.10/dist-packages (from pygwalker) (3.1.4)

Collecting kanaries-track==0.0.5 (from pygwalker)

Downloading kanaries_track-0.0.5-py3-none-any.whl (8.6 kB)

Requirement already satisfied: packaging in /usr/local/lib/python3.10/dist-packages (from pygwalker) (24.0)

Requirement already satisfied: pandas in /usr/local/lib/python3.10/dist-packages (from pygwalker) (2.0.3)

Requirement already satisfied: psutil in /usr/local/lib/python3.10/dist-packages (from pygwalker) (5.9.5)

Requirement already satisfied: pyarrow in /usr/local/lib/python3.10/dist-packages (from pygwalker) (14.0.2)

Requirement already satisfied: pydantic in /usr/local/lib/python3.10/dist-packages (from pygwalker) (2.7.1)

Requirement already satisfied: pytz in /usr/local/lib/python3.10/dist-packages (from pygwalker) (2023.4)

Requirement already satisfied: requests in /usr/local/lib/python3.10/dist-packages (from pygwalker) (2.31.0)

Collecting segment-analytics-python==2.2.3 (from pygwalker)

Downloading segment_analytics_python-2.2.3-py2.py3-none-any.whl (24 kB)

Requirement already satisfied: sqlalchemy in /usr/local/lib/python3.10/dist-packages (from pygwalker) (2.0.30)

Collecting sqlglot>=23.15.8 (from pygwalker)

Downloading sqlglot-24.0.1-py3-none-any.whl (374 kB)

374.7/374.7 kB

16.8 MB/s

eta 0:00:00

Requirement already satisfied: typing-extensions in /usr/local/lib/python3.10/dist-packages (from pygwalker) (4.11.0)

Collecting wasmtime>=12.0.0 (from gw-dsl-parser==0.1.48a6->pygwalker)

Downloading wasmtime-21.0.0-py3-none-manylinux1_x86_64.whl (5.4 MB)

5.4/5.4 MB

26.5 MB/s

eta 0:00:00

Collecting backoff>=2.2.1 (from kanaries-track==0.0.5->pygwalker)

Downloading backoff-2.2.1-py3-none-any.whl (15 kB)

Collecting dateutils>=0.6.12 (from kanaries-track==0.0.5->pygwalker)

Downloading dateutils-0.6.12-py2.py3-none-any.whl (5.7 kB)

Collecting monotonic~=1.5 (from segment-analytics-python==2.2.3->pygwalker)

Downloading monotonic-1.6-py2.py3-none-any.whl (8.2 kB)

Requirement already satisfied: python-dateutil~=2.2 in /usr/local/lib/python3.10/dist-packages (from segment-analytics-python==2.2.3->pygwalker)

Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.10/dist-packages (from requests->pygwalker) (3.3.2)

Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-packages (from requests->pygwalker) (3.7)

Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.10/dist-packages (from requests->pygwalker) (2.0.7)

Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.10/dist-packages (from requests->pygwalker) (2024.2.2)

Collecting types-python-dateutil>=2.8.10 (from arrow->pygwalker)

Downloading types_python_dateutil-2.9.0.20240316-py3-none-any.whl (9.7 kB)

Requirement already satisfied: setuptools>=18.5 in /usr/local/lib/python3.10/dist-packages (from ipython->pygwalker) (67.7.2)

Collecting jedi>=0.16 (from ipython->pygwalker)

Downloading jedi-0.19.1-py2.py3-none-any.whl (1.6 MB)

1.6/1.6 MB

40.4 MB/s

eta 0:00:00

Requirement already satisfied: decorator in /usr/local/lib/python3.10/dist-packages (from ipython->pygwalker) (4.4.2)

Requirement already satisfied: pickleshare in /usr/local/lib/python3.10/dist-packages (from ipython->pygwalker) (0.7.5)

Requirement already satisfied: traitlets>=4.2 in /usr/local/lib/python3.10/dist-packages (from ipython->pygwalker) (5.7.1)

```
import pandas as pd
import chardet
from sqlalchemy import create_engine
import matplotlib.pyplot as plt
from bokeh.plotting import figure, show, output_notebook
from bokeh.io import output_file
from bokeh.transform import factor_cmap
from bokeh.layouts import gridplot
from bokeh.palettes import Spectral6
from pygwalker import walk
```

Empieza a programar o a [crear código](#) con IA.

```
# Detectar la codificación del archivo
with open('/content/registrohistorico1.csv', 'rb') as f:
    result = chardet.detect(f.read(100000))
    encoding = result['encoding']
    print(f'Detected encoding: {encoding}')
```

Detected encoding: UTF-8-SIG

```
try:
    df = pd.read_csv('/content/registrohistorico1.csv',sep=';', on_bad_lines='skip',encoding=encoding)
except pd.errors.ParserError as e:
    print(f"ParserError: {e}")
```

<ipython-input-4-aa492f2ae796>:2: DtypeWarning: Columns (21,22,23,24,25,26) have mixed types. Specify dtype option on import or set low_memory=False
df = pd.read_csv('/content/registrohistorico1.csv',sep=';', on_bad_lines='skip',encoding=encoding)

df

	Anio_lectivo	Zona	Provincia	Cod_Provincia	Canton	Cod_Canton	Parroquia	Cod_Parroquia	Nombre_Institucion	AMIE	...	Total_I
0	2009-2010 Inicio	Zona 6	AZUAY	1	CUENCA	101	EL SAGRARIO	10104	UNIDAD EDUCATIVA PARTICULAR ROSA DE JESUS CORDERO	01B00002	...	
1	2009-2010 Inicio	Zona 6	AZUAY	1	CUENCA	101	MONAY	10109	CEBCI	01B00010	...	
2	2009-2010 Inicio	Zona 6	AZUAY	1	CUENCA	101	BAÑOS	10151	CENTRO EDUCATIVO ROUSSEAU	01B00019	...	
3	2009-2010 Inicio	Zona 6	AZUAY	1	CUENCA	101	BAÑOS	10151	COLEGIO INTERCULTURAL BILINGUE DE NARANCAY	01B00020	...	
4	2009-2010 Inicio	Zona 6	AZUAY	1	CUENCA	101	CHAUCHA	10153	SEIS DE JUNIO	01B00021	...	
...	
306445	2023-2024 Inicio	Zona 1	SUCUMBIOS	21	LAGO AGRIO	2101	SANTA CECILIA	210158	ZAMORA	21B00046	...	
306446	2023-2024 Inicio	Zona 1	CARCHI	4	MIRA	404	JIJON Y CAAMAÑO (CAB. EN RIO BLANCO)	40452	ZAMORA CHINCHIPE	04H00302	...	
306447	2023-2024 Inicio	Zona 9	PICHINCHA	17	QUITO	1701	PIFO	170175	ZAMORA CHINCHIPE	17H01913	...	
306448	2023-2024 Inicio	Zona 9	PICHINCHA	17	QUITO	1701	CALDERON (CARAPUNGO)	170155	ZARAN	17H01547	...	
306449	2023-2024 Inicio	Zona 4	MANABI	13	SAN VICENTE	1322	CANOA	132251	ZOILA CLEMENCIA CASTILLO	13H04341	...	

306450 rows × 27 columns

```
# Leer el archivo CSV
df2 = pd.read_csv('/content/registrohistorico2.csv',sep=';', on_bad_lines='skip',encoding=encoding)
```

df2



	Anio_lectivo	Zona	Provincia	Cod_Provincia	Canton	Cod_Canton	Parroquia	Cod_Parroquia	Nombre_Institucion	AMIE	...	area	R
0	2009-2010 Fin	Zona 6	AZUAY	1	CUENCA	101	BAÑOS	10151	CENTRO EDUCATIVO ROUSSEAU	01B00019	...	Rural	
1	2009-2010 Fin	Zona 6	AZUAY	1	CUENCA	101	BAÑOS	10151	COLEGIO INTERCULTURAL BILINGUE DE NARANCAY	01B00020	...	Rural	
2	2009-2010 Fin	Zona 6	AZUAY	1	CUENCA	101	CHAUCHA	10153	SEIS DE JUNIO	01B00021	...	Rural	
3	2009-2010 Fin	Zona 6	AZUAY	1	CUENCA	101	GIL RAMIREZ DAVALOS	10106	UNIDAD EDUCATIVA INTERCULTURAL BILINGUE LA PAZ...	01B00022	...	Urbana	
4	2009-2010 Fin	Zona 6	AZUAY	1	CUENCA	101	MOLLETURO	10157	LA PAZ	01B00023	...	Rural	
...	
295147	2022-2023 Fin	Zona 5	SANTA ELENA	24	SANTA ELENA	2401	SIMON BOLIVAR (JULIO MORENO)	240155	ESCUELA DE EDUCACION BASICA CARLOS ALBERTO FLORES	24H00423	...	Rural	
295148	2022-2023 Fin	Zona 5	SANTA ELENA	24	SALINAS	2403	CARLOS ESPINOZA LARREA	240301	UNIDAD EDUCATIVA CAP RAFAEL MORAN VALVERDE	24H00523	...	Urbana	
295149	2022-2023 Fin	Zona 5	SANTA ELENA	24	LA LIBERTAD	2402	LA LIBERTAD	240250	ESCUELA DE EDUCACIÓN BÁSICA JEAN PIAGET	24H00524	...	Urbana	
295150	2022-2023 Fin	Zona 5	SANTA ELENA	24	SALINAS	2403	JOSE LUIS TAMAYO (MUEY)	240352	ESCUELA DE EDUCACIÓN BÁSICA MONTESSORI	24H00525	...	Rural	
295151	2022-2023 Fin	Zona 5	SANTA ELENA	24	LA LIBERTAD	2402	LA LIBERTAD	240250	UNIDAD EDUCATIVA PCEI JAMES SMITHSON	24H00526	...	Urbana	

295152 rows × 23 columns

```
# Crear una conexión a una base de datos SQLite en un archivo
base_datos = create_engine('sqlite:///registro_historico2.db')

# Guardar el DataFrame en la base de datos SQL
df2.to_sql('registros_educacion', base_datos, if_exists='replace', index=False)
```



295152

```
#Leer los datos de la base de datos creada previamente

df3 = pd.read_sql('SELECT * FROM registros_educacion', base_datos)

# Mostrar los registros del DataFrame
df3
```



	Anio_lectivo	Zona	Provincia	Cod_Provincia	Canton	Cod_Canton	Parroquia	Cod_Parroquia	Nombre_Institucion	
0	2009-2010 Fin	Zona 6	AZUAY	1	CUENCA	101	BAÑOS	10151	CENTRO EDUCATIVO ROUSSEAU	01B0
1	2009-2010 Fin	Zona 6	AZUAY	1	CUENCA	101	BAÑOS	10151	COLEGIO INTERCULTURAL BILINGUE DE NARANCAY	01B0
2	2009-2010 Fin	Zona 6	AZUAY	1	CUENCA	101	CHAUCHA	10153	SEIS DE JUNIO	01B0
3	2009-2010 Fin	Zona 6	AZUAY	1	CUENCA	101	GIL RAMIREZ DAVALOS	10106	UNIDAD EDUCATIVA INTERCULTURAL BILINGUE LA PAZ...	01B0
4	2009-2010 Fin	Zona 6	AZUAY	1	CUENCA	101	MOLLETURO	10157	LA PAZ	01B0
...
295147	2022-2023 Fin	Zona 5	SANTA ELENA	24	SANTA ELENA	2401	SIMON BOLIVAR (JULIO MORENO)	240155	ESCUELA DE EDUCACION BASICA CARLOS ALBERTO FLORES	24H0
295148	2022-2023 Fin	Zona 5	SANTA ELENA	24	SALINAS	2403	CARLOS ESPINOZA LARREA	240301	UNIDAD EDUCATIVA CAP RAFAEL MORAN VALVERDE	24H0
295149	2022-2023 Fin	Zona 5	SANTA ELENA	24	LA LIBERTAD	2402	LA LIBERTAD	240250	ESCUELA DE EDUCACIÓN BÁSICA JEAN PIAGET	24H0
295150	2022-2023 Fin	Zona 5	SANTA ELENA	24	SALINAS	2403	JOSE LUIS TAMAYO (MUEY)	240352	ESCUELA DE EDUCACIÓN BÁSICA MONTESSORI	24H0
295151	2022-2023 Fin	Zona 5	SANTA ELENA	24	LA LIBERTAD	2402	LA LIBERTAD	240250	UNIDAD EDUCATIVA PCEI JAMES SMITHSON	24H0

295152 rows × 23 columns

```
# Cierra la conexión
conn.close()
```



```
-----
NameError                                Traceback (most recent call last)
<ipython-input-15-7b944705459a> in <cell line: 2>()
      1 # Cierra la conexión
----> 2 conn.close()

NameError: name 'conn' is not defined
```

```
df.describe()
```



	Cod_Provincia	Cod_Canton	Cod_Parroquia	Docentes_Femenino	Docentes_Masculino	Total_Docentes	Estudiantes_Femenino	Estudiantes_Masculi
count	306450.000000	306450.000000	306450.000000	306450.000000	306450.000000	306450.000000	306450.000000	306450.0000
mean	11.672230	1172.015285	117239.604229	7.545711	3.253320	10.799031	107.090419	108.9457
std	6.696917	669.239687	66924.927682	12.464926	6.590009	17.924272	217.463046	210.4353
min	1.000000	101.000000	10101.000000	0.000000	0.000000	0.000000	0.000000	0.0000
25%	8.000000	807.000000	80750.000000	1.000000	0.000000	1.000000	10.000000	11.0000
50%	11.000000	1112.000000	111250.000000	3.000000	1.000000	4.000000	28.000000	29.0000
75%	15.000000	1503.000000	150350.000000	9.000000	3.000000	12.000000	105.000000	110.0000
max	90.000000	9004.000000	900451.000000	251.000000	212.000000	374.000000	7543.000000	5880.0000

```
df.info()
```



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 306450 entries, 0 to 306449
Data columns (total 27 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Anio_lectivo           306450 non-null object
1   Zona                  306450 non-null object
2   Provincia              306450 non-null object
3   Cod_Provincia          306450 non-null int64
4   Canton                 306450 non-null object
5   Cod_Canton             306450 non-null int64
```

```
6 Parroquia 306450 non-null object
7 Cod_Parroquia 306450 non-null int64
8 Nombre_Institucion 306450 non-null object
9 AMIE 306450 non-null object
10 Tipo_Educacion 306450 non-null object
11 Sostenimiento 306450 non-null object
12 Area 306450 non-null object
13 Regimen_Escolar 306450 non-null object
14 Jurisdiccion 306450 non-null object
15 Docentes_Femenino 306450 non-null int64
16 Docentes_Masculino 306450 non-null int64
17 Total_Docentes 306450 non-null int64
18 Estudiantes_Femenino 306450 non-null int64
19 Estudiantes_Masculino 306450 non-null int64
20 Total_Estudiantes 306450 non-null float64
21 Ecuatoriana 306450 non-null object
22 Colombiana 306450 non-null object
23 Venezolana 306450 non-null object
24 Peruana 306450 non-null object
25 Otros_Paises_de_America 306450 non-null object
26 Otros_Continentes 306450 non-null object
dtypes: float64(1), int64(8), object(18)
memory usage: 63.1+ MB
```

df3.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 295152 entries, 0 to 295151
Data columns (total 23 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Anio_lectivo           295152 non-null object
1   Zona                  295152 non-null object
2   Provincia              295152 non-null object
3   Cod_Provincia          295152 non-null int64
4   Canton                 295152 non-null object
5   Cod_Canton             295152 non-null int64
6   Parroquia              295152 non-null object
7   Cod_Parroquia          295152 non-null int64
8   Nombre_Institucion     295152 non-null object
9   AMIE                   295152 non-null object
10  Escolarizacion         295152 non-null object
11  Tipo_Educacion         295152 non-null object
12  Sostenimiento          295152 non-null object
13  area                   295152 non-null object
14  Regimen_Escolar        295152 non-null object
15  Jurisdiccion           295152 non-null object
16  Modalidad              295152 non-null object
17  Jornada                295152 non-null object
18  Acceso_Edificio        295152 non-null object
19  Total_Estudiantes      295152 non-null int64
20  Promovidos             295152 non-null int64
21  No promovidos          295152 non-null int64
22  Abandono               295152 non-null int64
dtypes: int64(7), object(16)
memory usage: 51.8+ MB
```

```
# Dividir la columna 'Anio_lectivo' en dos partes utilizando un espacio como separador
# Esto se realiza para eliminar la palabra fin del dataframe y solo quedarnos con el año lectivo que servirá para
# hacer el merge con año lectivo y AMIE, que son las claves primarias de cada dataset
```

```
df['Anio_lectivo'] = df['Anio_lectivo'].str.split().str[0]
```

df



	Anio_lectivo	Zona	Provincia	Cod_Provincia	Canton	Cod_Canton	Parroquia	Cod_Parroquia	Nombre_Institucion	
0	2009-2010	Zona 6	AZUAY	1	CUENCA	101	EL SAGRARIO	10104	UNIDAD EDUCATIVA PARTICULAR ROSA DE JESUS CORDERO	01
1	2009-2010	Zona 6	AZUAY	1	CUENCA	101	MONAY	10109	CEBCI	01
2	2009-2010	Zona 6	AZUAY	1	CUENCA	101	BAÑOS	10151	CENTRO EDUCATIVO ROUSSEAU	01
3	2009-2010	Zona 6	AZUAY	1	CUENCA	101	BAÑOS	10151	COLEGIO INTERCULTURAL BILINGUE DE NARANCAY	01
4	2009-2010	Zona 6	AZUAY	1	CUENCA	101	CHAUCHA	10153	SEIS DE JUNIO	01
...
306445	2023-2024	Zona 1	SUCUMBIOS	21	LAGO AGRIO	2101	SANTA CECILIA	210158	ZAMORA	21
306446	2023-2024	Zona 1	CARCHI	4	MIRA	404	JIJON Y CAAMAÑO (CAB. EN RIO BLANCO)	40452	ZAMORA CHINCHIPE	04
306447	2023-2024	Zona 9	PICHINCHA	17	QUITO	1701	PIFO	170175	ZAMORA CHINCHIPE	17
306448	2023-2024	Zona 9	PICHINCHA	17	QUITO	1701	CALDERON (CARAPUNGO)	170155	ZARAN	17
306449	2023-2024	Zona 4	MANABI	13	SAN VICENTE	1322	CANOA	132251	ZOILA CLEMENCIA CASTILLO	13

306450 rows × 27 columns

Dividir la columna 'Anio_lectivo' en dos partes utilizando un espacio como separador
Esto se realiza para eliminar la palabra fin del dataframe y solo quedarnos con el año lectivo que servirá para
hacer el merge con año lectivo y AMIE, que son las claves primarias de cada dataset

df3['Anio_lectivo'] = df['Anio_lectivo'].str.split().str[0]

df3



	Anio_lectivo	Zona	Provincia	Cod_Provincia	Canton	Cod_Canton	Parroquia	Cod_Parroquia	Nombre_Institucion	
0	2009-2010	Zona 6	AZUAY	1	CUENCA	101	BAÑOS	10151	CENTRO EDUCATIVO ROUSSEAU	01B0
1	2009-2010	Zona 6	AZUAY	1	CUENCA	101	BAÑOS	10151	COLEGIO INTERCULTURAL BILINGUE DE NARANCAY	01B0
2	2009-2010	Zona 6	AZUAY	1	CUENCA	101	CHAUCHA	10153	SEIS DE JUNIO	01B0
3	2009-2010	Zona 6	AZUAY	1	CUENCA	101	GIL RAMIREZ DAVALOS	10106	UNIDAD EDUCATIVA INTERCULTURAL BILINGUE LA PAZ...	01B0
4	2009-2010	Zona 6	AZUAY	1	CUENCA	101	MOLLETURO	10157	LA PAZ	01B0
...
295147	2023-2024	Zona 5	SANTA ELENA	24	SANTA ELENA	2401	SIMON BOLIVAR (JULIO MORENO)	240155	ESCUELA DE EDUCACION BASICA CARLOS ALBERTO FLORES	24H0
295148	2023-2024	Zona 5	SANTA ELENA	24	SALINAS	2403	CARLOS ESPINOZA LARREA	240301	UNIDAD EDUCATIVA CAP RAFAEL MORAN VALVERDE	24H0
295149	2023-2024	Zona 5	SANTA ELENA	24	LA LIBERTAD	2402	LA LIBERTAD	240250	ESCUELA DE EDUCACIÓN BÁSICA JEAN PIAGET	24H0
295150	2023-2024	Zona 5	SANTA ELENA	24	SALINAS	2403	JOSE LUIS TAMAYO (MUEY)	240352	ESCUELA DE EDUCACIÓN BÁSICA MONTESSORI	24H0
295151	2023-2024	Zona 5	SANTA ELENA	24	LA LIBERTAD	2402	LA LIBERTAD	240250	UNIDAD EDUCATIVA PCEI JAMES SMITHSON	24H0


295152 rows × 23 columns

Se agrega las columnas provenientes del dataframe (base de datos) al dataframe final (merged_df).

Realizar el merge en base a dos columnas y especificar qué columnas agregar del segundo dataset

merged_df = pd.merge(df, df3[['Anio_lectivo', 'AMIE', 'Modalidad','Jornada', 'Acceso_Edificio', 'Promovidos', 'No promovidos', 'Abandono']], on=['Anio_


merged_df



	Anio_lectivo	Zona	Provincia	Cod_Provincia	Canton	Cod_Canton	Parroquia	Cod_Parroquia	Nombre_Institucion
0	2009-2010	Zona 6	AZUAY	1	CUENCA	101	BAÑOS	10151	CENTRO EDUCATIVO ROUSSEAU
1	2009-2010	Zona 6	AZUAY	1	CUENCA	101	BAÑOS	10151	COLEGIO INTERCULTURAL BILINGUE DE NARANCAY
2	2009-2010	Zona 6	AZUAY	1	CUENCA	101	CHAUCHA	10153	SEIS DE JUNIO
3	2009-2010	Zona 6	AZUAY	1	CUENCA	101	GIL RAMIREZ DAVALOS	10106	LA PAZ INTERCULTURAL BILINGUE AZUAY
4	2009-2010	Zona 6	AZUAY	1	CUENCA	101	MOLLETURO	10157	LA PAZ
...
281648	2023-2024	Zona 2	NAPO	15	ARCHIDONA	1503	ARCHIDONA	150350	YAWARI
281649	2023-2024	Zona 9	PICHINCHA	17	QUITO	1701	YARUQUI	170185	YOLANDA MEDINA MENA
281650	2023-2024	Zona 1	SUCUMBIOS	21	LAGO AGRIO	2101	SANTA CECILIA	210158	ZAMORA
281651	2023-2024	Zona 9	PICHINCHA	17	QUITO	1701	PIFO	170175	ZAMORA CHINCHIPE
281652	2023-2024	Zona 9	PICHINCHA	17	QUITO	1701	CALDERON (CARAPUNGO)	170155	ZARAN

281653 rows × 33 columns

merged_df.info()



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 281653 entries, 0 to 281652
Data columns (total 33 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   Anio_lectivo                          281653 non-null object
1   Zona                                281653 non-null object
2   Provincia                             281653 non-null object
3   Cod_Provincia                         281653 non-null int64
4   Canton                               281653 non-null object
5   Cod_Canton                           281653 non-null int64
6   Parroquia                             281653 non-null object
7   Cod_Parroquia                         281653 non-null int64
8   Nombre_Institucion                   281653 non-null object
9   AMIE                                 281653 non-null object
10  Tipo_Educacion                        281653 non-null object
11  Sostenimiento                         281653 non-null object
12  Area                                  281653 non-null object
13  Regimen_Escolar                       281653 non-null object
14  Jurisdiccion                          281653 non-null object
15  Docentes_Femenino                     281653 non-null int64
16  Docentes_Masculino                     281653 non-null int64
17  Total_Docentes                        281653 non-null int64
18  Estudiantes_Femenino                  281653 non-null int64
19  Estudiantes_Masculino                  281653 non-null int64
20  Total_Estudiantes                     281653 non-null float64
21  Ecuatoriana                           281653 non-null object
22  Colombiana                            281653 non-null object
23  Venezolana                            281653 non-null object
24  Peruana                               281653 non-null object
25  Otros_Paises_de_America                281653 non-null object
26  Otros_Continentes                     281653 non-null object
27  Modalidad                             281653 non-null object
28  Jornada                               281653 non-null object
29  Acceso_Edificio                       281653 non-null object
30  Promovidos                            281653 non-null int64
31  No promovidos                         281653 non-null int64
32  Abandono                              281653 non-null int64
dtypes: float64(1), int64(11), object(21)
memory usage: 70.9+ MB
```

```
merged_df['Anio_lectivo'].value_counts()
```

Anio_lectivo	
2011-2012	25951
2012-2013	25100
2009-2010	24999
2010-2011	24295
2013-2014	23247
2014-2015	21980
2015-2016	17949
2016-2017	16685
2017-2018	16287
2018-2019	16276
2019-2020	16235
2020-2021	16011
2021-2022	15927
2022-2023	15912
2023-2024	4799
Name: count, dtype: int64	

Guardamos el df final a un archivo CSV

```
merged_df.to_csv('/content/registros_final.csv', index=False) # index=False evita que se agregue la columna de índices al archivo CSV
```

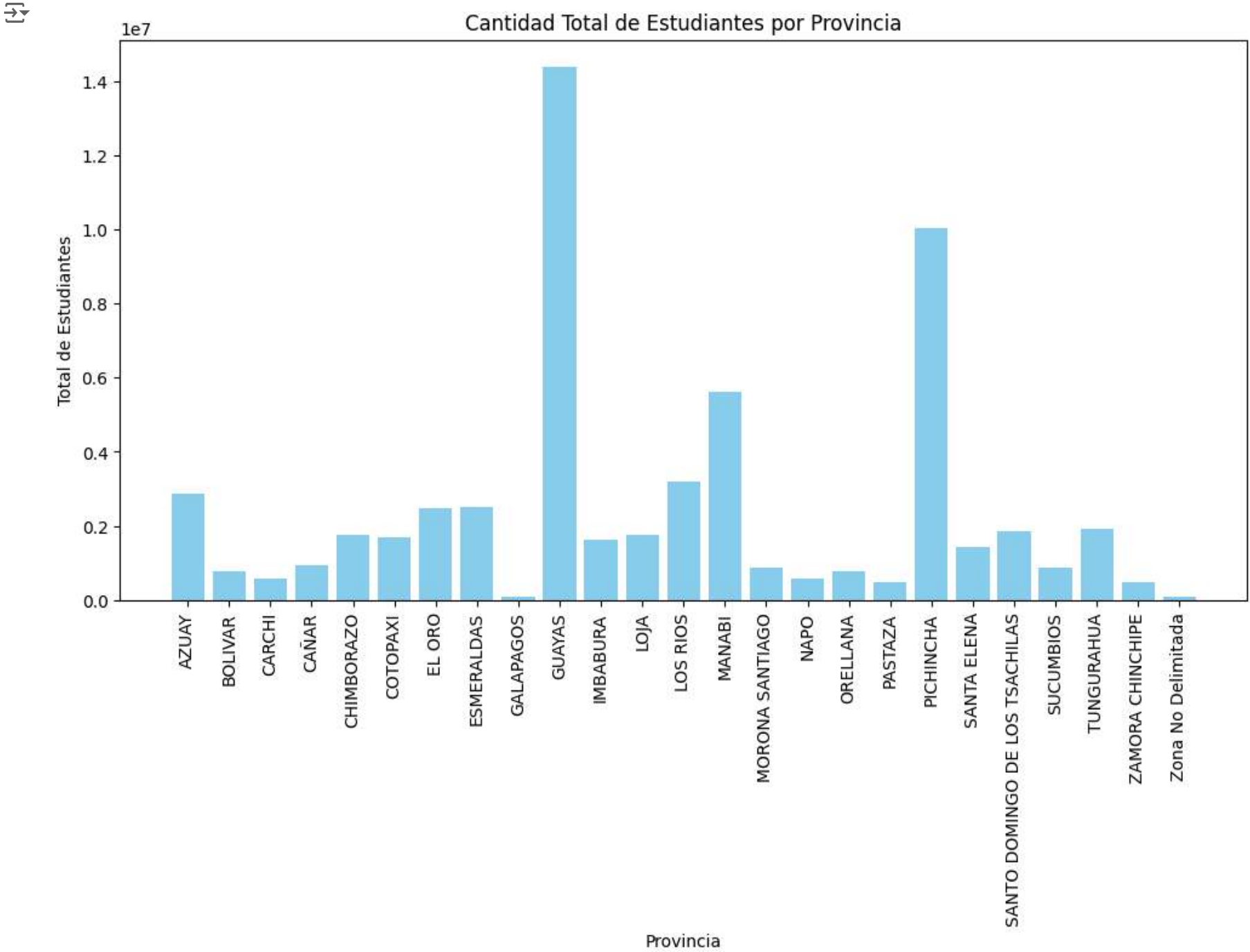
GRÁFICOS / VISUALIZACIONES

Matplotlib

Gráfico de barras de la cantidad total de estudiantes por provincia.

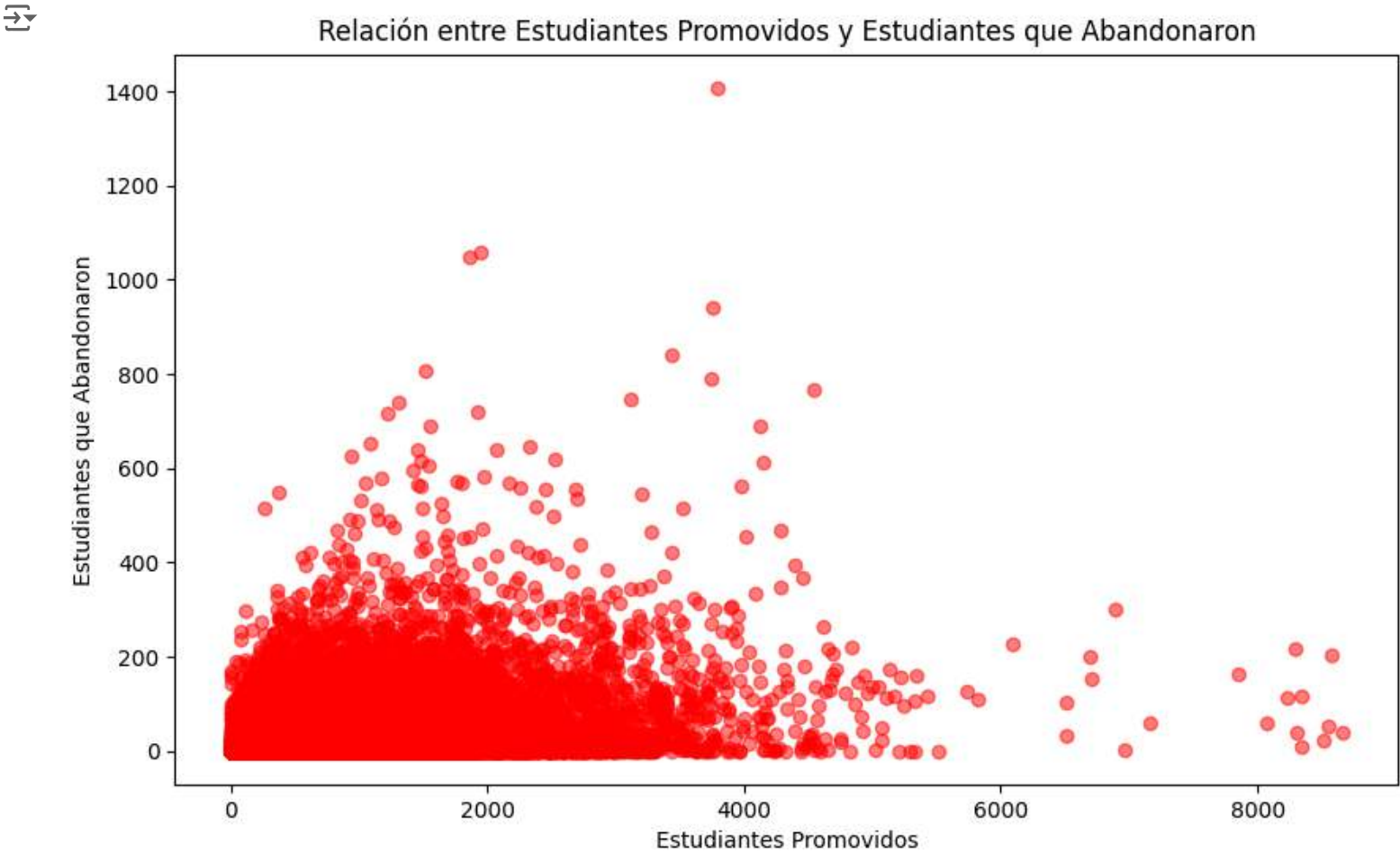
```
# Agrupar datos por provincia y sumar el total de estudiantes
grouped_data = merged_df.groupby('Provincia')['Total_Estudiantes'].sum().reset_index()

# Crear gráfico de barras
plt.figure(figsize=(12, 6))
plt.bar(grouped_data['Provincia'], grouped_data['Total_Estudiantes'], color='skyblue')
plt.xlabel('Provincia')
plt.ylabel('Total de Estudiantes')
plt.title('Cantidad Total de Estudiantes por Provincia')
plt.xticks(rotation=90)
plt.show()
```



Ahora, vamos a generar un diagrama de dispersi3n para visualizar la relaci3n entre estudiantes promovidos y estudiantes que abandonaron.

```
# Crear diagrama de dispersi3n
plt.figure(figsize=(10, 6))
plt.scatter(merged_df['Promovidos'], merged_df['Abandono'], alpha=0.5, c='red')
plt.xlabel('Estudiantes Promovidos')
plt.ylabel('Estudiantes que Abandonaron')
plt.title('Relaci3n entre Estudiantes Promovidos y Estudiantes que Abandonaron')
plt.show()
```



Aqu3 tenemos dos visualizaciones generadas con Matplotlib:

- Gr3fico de barras que muestra la cantidad total de estudiantes por provincia.
- Diagrama de dispersi3n que visualiza la relaci3n entre estudiantes promovidos y estudiantes que abandonaron.

✓ **Bokeh**

```
from bokeh.plotting import figure, show, output_file, output_notebook
from bokeh.models import ColumnDataSource
from bokeh.layouts import gridplot
from bokeh.palettes import Spectral11 # Asumiendo que necesitas más colores

# Agrupar por provincia y año lectivo y sumar las cantidades
df_grouped = merged_df.groupby(['Provincia', 'Anio_lectivo']).agg({'Total_Estudiantes': 'sum'}).reset_index()

# Convertir las columnas 'Provincia' y 'Año' a string para evitar problemas en Bokeh
df_grouped['Provincia'] = df_grouped['Provincia'].astype(str)
df_grouped['Anio_lectivo'] = df_grouped['Anio_lectivo'].astype(str)

# Preparar los datos para el gráfico de barras
provincias = df_grouped['Provincia'].unique().tolist()
years = df_grouped['Anio_lectivo'].unique().tolist()

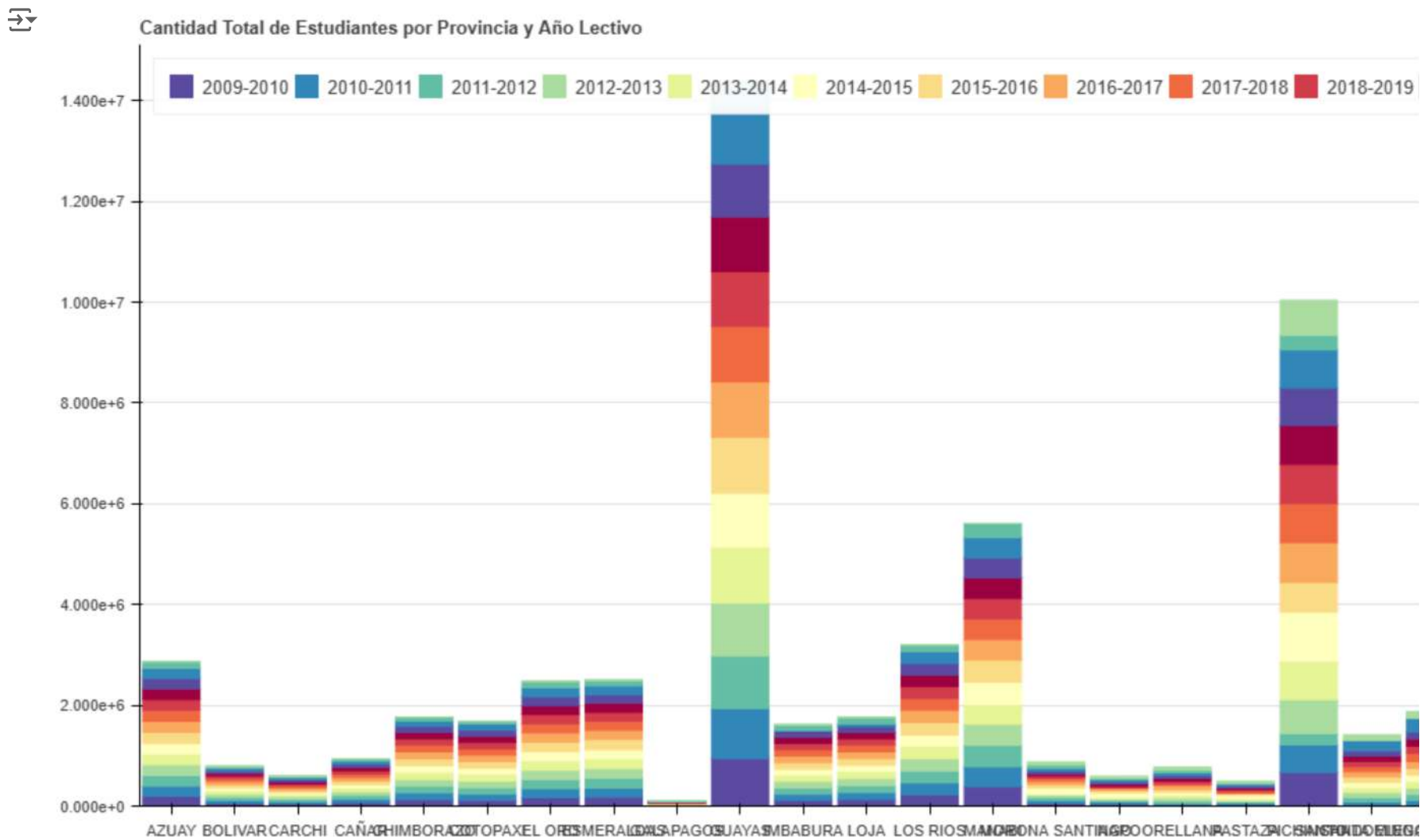
# Asegurarte de que la paleta tenga suficientes colores para los años, repitiendo colores si es necesario
if len(years) > 11:
    color_palette = [Spectral11[i % 11] for i in range(len(years))]
else:
    color_palette = Spectral11[:len(years)]

# Crear un diccionario para almacenar los datos
data = {'Provincia': provincias}
for year in years:
    data[year] = [df_grouped[(df_grouped['Provincia'] == provincia) & (df_grouped['Anio_lectivo'] == year)]['Total_Estudiantes'].sum() for provincia in provincias]

# Crear el gráfico de barras apiladas
source = ColumnDataSource(data)

p1 = figure(x_range=provincias, title="Cantidad Total de Estudiantes por Provincia y Año Lectivo",
            toolbar_location=None, tools="", width=1200, height=600)
p1.vbar_stack(years, x='Provincia', width=0.9, color=color_palette, source=source,
            legend_label=years)

p1.y_range.start = 0
p1.xgrid.grid_line_color = None
p1.axis.minor_tick_line_color = None
p1.outline_line_color = None
p1.legend.location = "top_left"
p1.legend.orientation = "horizontal"
show(p1)
```



```
from bokeh.io import show, output_notebook
from bokeh.plotting import figure
from bokeh.transform import dodge
from bokeh.models import ColumnDataSource, LinearColorMapper, ColorBar
from bokeh.transform import transform
from bokeh.layouts import column

# Convertir columnas relevantes a numéricas, manejando errores
df['Promovidos'] = pd.to_numeric(merged_df['Promovidos'], errors='coerce').fillna(0).astype(int)
df['No promovidos'] = pd.to_numeric(merged_df['No promovidos'], errors='coerce').fillna(0).astype(int)
df['Abandono'] = pd.to_numeric(merged_df['Abandono'], errors='coerce').fillna(0).astype(int)

# Agrupar los datos por provincia y sumar las columnas relevantes
grouped = merged_df.groupby('Provincia')[['Promovidos', 'No promovidos', 'Abandono']].sum().reset_index()

# Crear la fuente de datos para Bokeh
source = ColumnDataSource(grouped)

output_notebook()

# Configuración del gráfico de barras apiladas
p1 = figure(x_range=grouped['Provincia'], height=400, title="Estudiantes Promovidos, No Promovidos y Abandonos por Provincia",
            toolbar_location=None, tools="")

p1.vbar(x=dodge('Provincia', -0.25, range=p1.x_range), top='Promovidos', width=0.2, source=source,
        color="green", legend_label="Promovidos")

p1.vbar(x=dodge('Provincia', 0.0, range=p1.x_range), top='No promovidos', width=0.2, source=source,
        color="red", legend_label="No Promovidos")

p1.vbar(x=dodge('Provincia', 0.25, range=p1.x_range), top='Abandono', width=0.2, source=source,
        color="blue", legend_label="Abandono")

p1.xaxis.major_label_orientation = 1.2
p1.xgrid.grid_line_color = None
p1.legend.title = 'Estado'
p1.legend.title_text_font_style = 'bold'
p1.legend.title_text_font_size = '10pt'
p1.legend.label_text_font_size = '8pt'
p1.legend.orientation = "horizontal"
p1.legend.location = "top_center"

# Mostrar la gráfica
show(p1)
```



Pygwalker

- Gráfica de abandono de estudiantes por año lectivo

```
# Iniciar la exploración de datos con Pywalker
walk(merged_df)
```

Loading Graphic-Walker UI...