Tc. Programación en Analítica de datos 228117

1901119 APLICAR BUENAS PRACTICAS PARA PREPARAR, LIMPIAR, REFINAR Y EXPLORAR GRANDES VOLUMENES DE DATOS EN EL SECTOR PRODUCTIVO.

NCL ORGANIZAR LA INFORMACIÓN A GESTIONAR DE ACUERDO CON TÉCNICAS DE ANÁLISIS.

NCL PROCESO DE DATOS DE ACUERDO CON PROCEDIMIENTO TÉCNICO Y METODOLOGÍA ESTADÍSTICA
RAP 45 ORGANIZAR LA INFORMACIÓN A GESTIONAR DE ACUERDO CON TÉCNICAS DE ANÁLISIS.

RAP 46 ELABORAR INFORMES UTILIZANDO HERRAMIENTA INFORMÁTICA SELECCIONADA.

RAP 50 RECOLECTAR INFORMACIÓN DE ACUERDO A LAS NECESIDADES DEL CLIENTE.

RAP 51 ORGANIZAR LA MUESTRA DE DATOS DE ACUERDO A LAS METODOLOGÍAS ESTADÍSTICAS.

RAP 52 REALIZAR PROCEDIMIENTOS SOBRE LOS DATOS APLICANDO VARIABLES Y TÉCNICAS ESTADÍSTICAS.

RAP 49 ELABORAR INFORMES SEGÚN LA NECESIDAD DEL CLIENTE

Carga de datos

Instructor: José Fernando Galindo Suarez
jgalindos@sena.edu.co
CGMLTI 2023

LIBRERIA NUMPY

CONTENIDO

- Cargar archivos CSV
- Cargar archivos Excel
- Cargar desde una URL

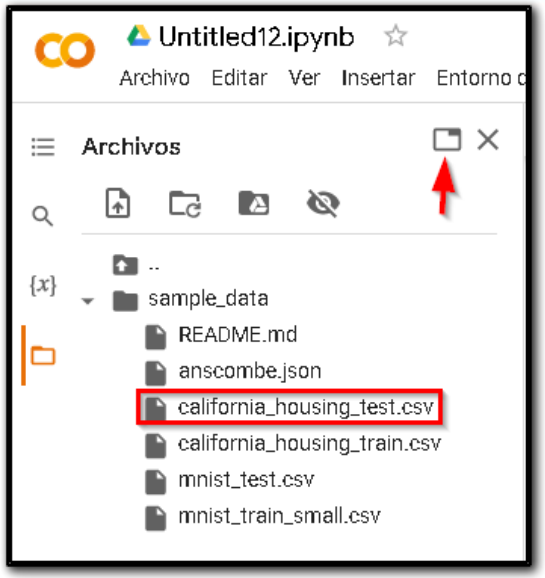
JOSE FERNANDO GALINDO SUAREZ

PANDAS.READ_CSV

pandas.read_csv(filepath_or_buffer, *, sep=_NoDefault.no_default, delimiter=None,
header='infer',
names=_NoDefault.no_default, index_col=None, usecols=None, squeeze=None,
prefix=_NoDefault.no_default,
mangle_dupe_cols=True, dtype=None, engine=None, converters=None, true_values=None,
false_values=None,
skipinitialspace=False, skiprows=None, skipfooter=0, nrows=None, na_values=None,
keep_default_na=True,
na_filter=True, verbose=False, skip_blank_lines=True, parse_dates=None,
infer_datetime_format=False,
keep_date_col=False, date_parser=None, dayfirst=False, cache_dates=True, iterator=False,
chunksize=None,
compression='infer', thousands=None, decimal='.', lineterminator=None, quotechar='"',
quoting=0, doublequote=True, escapechar=None, comment=None,
encoding=None, encoding_errors='strict', dialect=None, error_bad_lines=None,
warn_bad_lines=None, on_bad_lines=None, delim_whitespace=False, low_memory=True,
memory_map=False, float_precision=None, storage_options=None)

JOSE FERNANDO GALINDO SUAREZ

PANDAS.READ_CSV




```
import pandas as pd
import numpy as np
df=pd.read_csv("/content/sample_data/california_housing_test.csv")
df.describe()
```

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value
count	3000.000000	3000.00000	3000.000000	3000.000000	3000.000000	3000.000000	3000.00000	3000.000000	3000.00000
mean	-119.589200	35.63539	28.845333	2599.578667	529.950667	1402.798667	489.91200	3.807272	205846.27500
std	1.994936	2.12967	12.555396	2155.593332	415.654368	1030.543012	365.42271	1.854512	113119.68747
min	-124.180000	32.56000	1.000000	6.000000	2.000000	5.000000	2.00000	0.499900	22500.00000
25%	-121.810000	33.93000	18.000000	1401.000000	291.000000	780.000000	273.00000	2.544000	121200.00000
50%	-118.485000	34.27000	29.000000	2106.000000	437.000000	1155.000000	409.50000	3.487150	177650.00000
75%	-118.020000	37.69000	37.000000	3129.000000	636.000000	1742.750000	597.25000	4.656475	263975.00000
max	-114.490000	41.92000	52.000000	30450.000000	5419.000000	11935.000000	4930.00000	15.000100	500001.00000

JOSE FERNANDO GALINDO SUAREZ

PANDAS.READ_CSV

```
df.median() # Mediana
```

```
longitude           -118.48500
latitude              34.27000
housing_median_age    29.00000
total_rooms         2106.00000
total_bedrooms       437.00000
population          1155.00000
households           409.50000
median_income          3.48715
median_house_value  177650.00000
dtype: float64
```

JOSE FERNANDO GALINDO SUAREZ

PANDAS.READ_CSV

JOSE FERNANDO GALINDO SUAREZ

```
df.mode() # Moda
```

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value
0	-118.26	34.02	52.0	907.0	314.0	870.0	273.0	15.0001	500001.0
1	-118.21	NaN	NaN	1778.0	NaN	NaN	375.0	NaN	NaN
2	NaN	NaN	NaN	1787.0	NaN	NaN	614.0	NaN	NaN
3	NaN	NaN	NaN	1966.0	NaN	NaN	NaN	NaN	NaN

PANDAS.READ_CSV

```
df.groupby("population").mean() ## Calculo de la media
```

1

population	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	households	median_income	median_house_value
5.0	-114.620	33.62	26.0	18.0	3.0	3.0	0.53600	275000.0
8.0	-117.035	33.26	26.5	11.0	3.0	2.5	1.37500	57500.0
14.0	-117.775	33.35	50.5	28.5	6.0	8.0	1.90625	171900.0
19.0	-122.490	38.00	26.0	48.0	8.0	8.0	7.71970	400000.0
21.0	-118.060	34.03	36.0	21.0	7.0	9.0	2.37500	175000.0
...	...	...	...	...	...	...	...	...
8824.0	-117.120	33.49	4.0	21988.0	4055.0	3252.0	3.99630	191100.0
9419.0	-117.200	33.58	2.0	30450.0	5033.0	3197.0	4.59360	174300.0
10877.0	-117.270	33.15	4.0	23915.0	4135.0	3958.0	4.63570	244900.0
11139.0	-116.140	34.45	12.0	8796.0	1721.0	1680.0	2.26120	137500.0
11935.0	-121.530	38.48	5.0	27870.0	5027.0	4855.0	4.88110	212200.0

1802 rows × 8 columns

2

JOSE FERNANDO GALINDO SUAREZ

PANDAS.READ_CSV

```
df.groupby(["population","total_rooms"]).std() ##Calculo de la desviación estándar
```

		longitude	latitude	housing_median_age	total_bedrooms	households	median_income	median_house_value
population	**total_rooms**							
5.0	**18.0**	NaN	NaN	NaN	NaN	NaN	NaN	NaN
8.0	**6.0**	NaN	NaN	NaN	NaN	NaN	NaN	NaN
	16.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
14.0	**25.0**	NaN	NaN	NaN	NaN	NaN	NaN	NaN
	32.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
...	**...**	...	...	...	...	...	...	...
8824.0	**21988.0**	NaN	NaN	NaN	NaN	NaN	NaN	NaN
9419.0	**30450.0**	NaN	NaN	NaN	NaN	NaN	NaN	NaN
10877.0	**23915.0**	NaN	NaN	NaN	NaN	NaN	NaN	NaN
11139.0	**8796.0**	NaN	NaN	NaN	NaN	NaN	NaN	NaN
11935.0	**27870.0**	NaN	NaN	NaN	NaN	NaN	NaN	NaN

3000 rows × 7 columns

JOSE FERNANDO GALINDO SUAREZ

PANDAS.READ_CSV

```
#Cálculo de dos medidas estadísticas
df.groupby(["population","total_rooms"]).agg(['mean', 'std'])
```

		longitude		latitude		housing_median_age		total_bedrooms		households		median_income		median_house_value	
		mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
population	**total_rooms**														
5.0	**18.0**	-114.62	NaN	33.62	NaN	26.0	NaN	3.0	NaN	3.0	NaN	0.5360	NaN	275000.0	NaN
8.0	**6.0**	-116.95	NaN	33.86	NaN	1.0	NaN	2.0	NaN	2.0	NaN	1.6250	NaN	55000.0	NaN
	16.0	-117.12	NaN	32.66	NaN	52.0	NaN	4.0	NaN	3.0	NaN	1.1250	NaN	60000.0	NaN
14.0	**25.0**	-117.11	NaN	32.66	NaN	52.0	NaN	5.0	NaN	9.0	NaN	1.6250	NaN	118800.0	NaN
	32.0	-118.44	NaN	34.04	NaN	49.0	NaN	7.0	NaN	7.0	NaN	2.1875	NaN	225000.0	NaN
...	**...**	...	...	...	...	...	...	...	...	...	...	...	...	...	...
8824.0	**21988.0**	-117.12	NaN	33.49	NaN	4.0	NaN	4055.0	NaN	3252.0	NaN	3.9963	NaN	191100.0	NaN
9419.0	**30450.0**	-117.20	NaN	33.58	NaN	2.0	NaN	5033.0	NaN	3197.0	NaN	4.5936	NaN	174300.0	NaN
10877.0	**23915.0**	-117.27	NaN	33.15	NaN	4.0	NaN	4135.0	NaN	3958.0	NaN	4.6357	NaN	244900.0	NaN
11139.0	**8796.0**	-116.14	NaN	34.45	NaN	12.0	NaN	1721.0	NaN	1680.0	NaN	2.2612	NaN	137500.0	NaN
11935.0	**27870.0**	-121.53	NaN	38.48	NaN	5.0	NaN	5027.0	NaN	4855.0	NaN	4.8811	NaN	212200.0	NaN

3000 rows × 14 columns

JOSE FERNANDO GALINDO SUAREZ

PANDAS.READ_CSV

```
# El commando describe presenta la cantidad de datos de la variable, la cantidad de valores distintos, la moda y la frecuencia.
df["total_rooms"].describe()
```

```
count     3000.000000
mean      2599.578667
std       2155.593332
min          6.000000
25%       1401.000000
50%       2106.000000
75%       3129.000000
max      30450.000000
Name: total_rooms, dtype: float64
```

JOSE FERNANDO GALINDO SUAREZ

PANDAS.READ_CSV

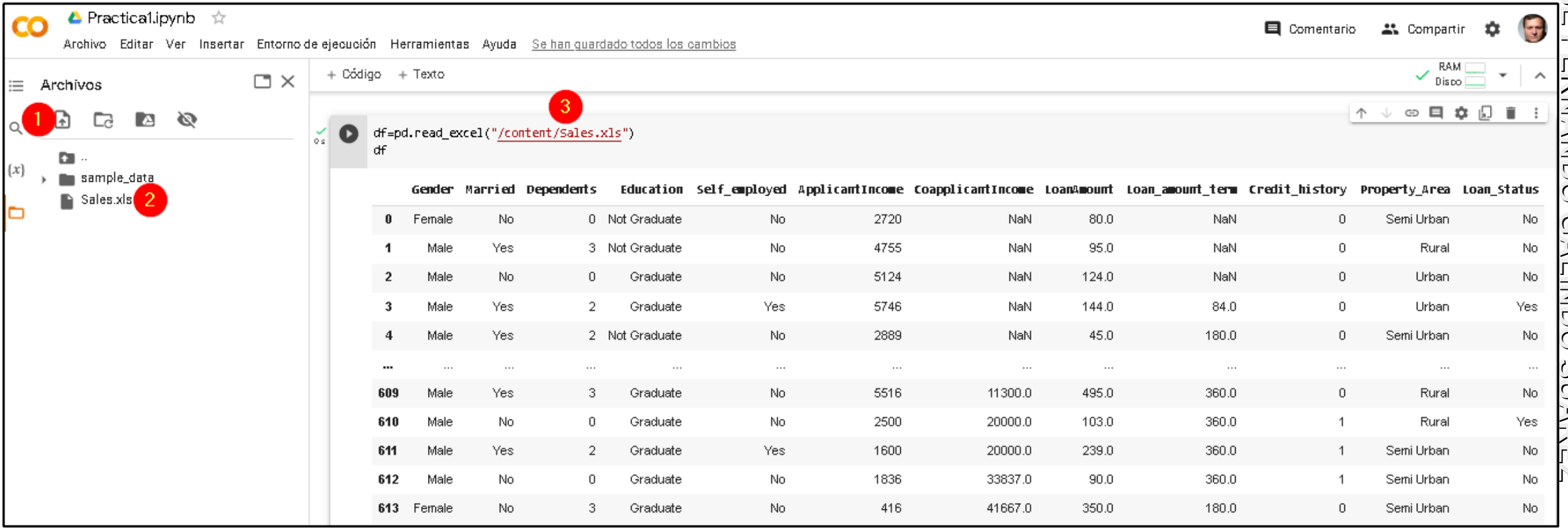
```
ts = pd.DataFrame(df.total_rooms) ## Se define el vector total_rooms
ts
```

	total_rooms
0	3885.0
1	1510.0
2	3589.0
3	67.0
4	1241.0
...	...
2995	1450.0
2996	5257.0
2997	956.0
2998	96.0
2999	1765.0

3000 rows × 1 columns

JOSE FERNANDO GALINDO SUAREZ

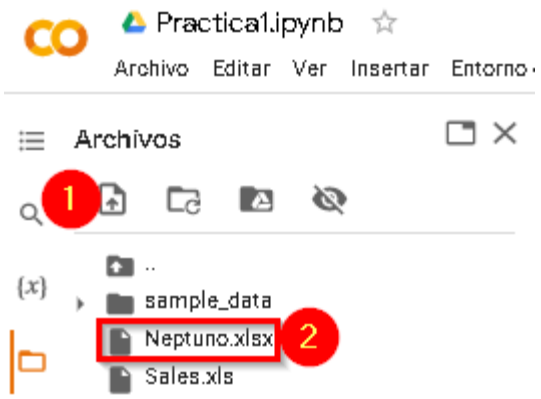
PANDAS.READ_CSV



	Gender	Married	Dependents	Education	Self_employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_amount_term	Credit_history	Property_Area	Loan_Status
0	Female	No	0	Not Graduate	No	2720	NaN	80.0	NaN	0	Semi Urban	No
1	Male	Yes	3	Not Graduate	No	4755	NaN	95.0	NaN	0	Rural	No
2	Male	No	0	Graduate	No	5124	NaN	124.0	NaN	0	Urban	No
3	Male	Yes	2	Graduate	Yes	5746	NaN	144.0	84.0	0	Urban	Yes
4	Male	Yes	2	Not Graduate	No	2889	NaN	45.0	180.0	0	Semi Urban	No
...	...	...	...	...	...	...	...	...	...	...	...	...
609	Male	Yes	3	Graduate	No	5516	11300.0	495.0	360.0	0	Rural	No
610	Male	No	0	Graduate	No	2500	20000.0	103.0	360.0	1	Rural	Yes
611	Male	Yes	2	Graduate	Yes	1600	20000.0	239.0	360.0	1	Semi Urban	No
612	Male	No	0	Graduate	No	1836	33837.0	90.0	360.0	1	Semi Urban	No
613	Female	No	3	Graduate	No	416	41667.0	350.0	180.0	0	Semi Urban	No

JOSE FERNANDO GALINDO SUAREZ

PANDAS.READ_CSV




```
df=pd.read_excel("/content/Neptuno.xlsx",sheet_name="2,-Categorias")
df
```

	Idcategoria	Nombrecategoria	Descripcion
0	1	Bebidas	Gaseosas, café, té, cervezas y maltas
1	2	Condimentos	Salsas dulces y picantes, delicias, comida par...
2	3	Repostería	Postres, dulces y pan dulce
3	4	Lácteos	Quesos
4	5	Granos/Cereales	Pan, galletas, pasta y cereales
5	6	Carnes	Carnes preparadas
6	7	Frutas/Verduras	Frutas secas y queso de soja
7	8	Pescado/Marisco	Pescados, mariscos y algas

JOSE FERNANDO GALINDO SUAREZ

PANDAS.READ_CSV

```
df=pd.read_csv("https://www.datos.gov.co/resource/tyhr-3h8y.csv")
df
```

	comuna	positivos_confirmados	pc	recuperados	r	activos	a
0	Comuna 1	619	1	482	0.78	137	0.22
1	CENTRO	229	1	182	0.79	47	0.21
2	OBRERO	72	1	54	0.75	18	0.25
3	SANTIAGO	59	1	42	0.71	17	0.29
4	LAS CUADRAS	54	1	43	0.80	11	0.20
...	...	...	...	...	...	...	...
386	VILLA ADRIANA MARIA	1	1	1	1.00	0	0.00
387	VILLA DE LOS ANDES	1	1	1	1.00	0	0.00
388	VILLA DEL PRADO	1	1	0	0.00	1	1.00
389	VILLA RECREO III	1	1	0	0.00	1	1.00
390	VILLAS DEL VIENTO	1	1	1	1.00	0	0.00

391 rows × 7 columns

Fuente: https://dev.socrata.com/foundry/www.datos.gov.co/tyhr-3h8y

JOSE FERNANDO GALINDO SUAREZ

PANDAS.READ_CSV

```
import json
df = pd.read_json("https://www.datos.gov.co/resource/52mm-fccv.json")
df
```

	institucion_sede_tipo_discapacidad_genero	baja_vision_irreversible_f_	baja_vision_irreversible_m_	ceguera_f_	ceguera_m_	espectro_autista_m_	_intelectual_f_	_intelectual_m_	multiple_f_	multiple_m_
0	Institucion_ genero	F	M	F	M	M	F	M	F	M
1	AGUSTIN CODAZZI	S/D	1	S/D	S/D	S/D	S/D	8	26	1
2	Agustin Codazzi	S/D	1	S/D	S/D	S/D	S/D	7	20	1
3	El Rosario	S/D	S/D	S/D	S/D	S/D	S/D	1	5	S/D
4	Emaya	S/D	S/D	S/D	S/D	S/D	S/D	S/D	1	S/D
...	...	...	...	...	...	...	...	...	...	...
123	Tecnico I.P.C. Andres Rosa	1	S/D	S/D	S/D	S/D	S/D	1	6	S/D
124	TECNICO SUPERIOR	S/D	S/D	S/D	S/D	S/D	S/D	10	5	1
125	Los Martires	S/D	S/D	S/D	S/D	S/D	S/D	S/D	S/D	S/D
126	Tecnico Superior	S/D	S/D	S/D	S/D	S/D	S/D	10	5	1
127	Total general	8	14	6	6	2	14	275	481	20

128 rows × 22 columns

Fuente: https://dev.socrata.com/foundry/www.datos.gov.co/52mm-fccv

JOSE FERNANDO GALINDO SUAREZ

PANDAS.READ_CSV

```
import json
df = pd.read_json("/content/incapacidad.json")
df
```

	institucion_sede_tipo_discapacidad_genero	baja_vision_irreversible_f_	baja_vision_irreversible_m_	ceguera_f_	ceguera_m_	espectro_autista_m_	_intelectual_f_	_intelectual_m_	multiple_f_	multiple_m_
0	Institucion_ genero	F	M	F	M	M	F	M	F	M
1	AGUSTIN CODAZZI	S/D	1	S/D	S/D	S/D	S/D	8	26	1
2	Agustin Codazzi	S/D	1	S/D	S/D	S/D	S/D	7	20	1
3	El Rosario	S/D	S/D	S/D	S/D	S/D	S/D	1	5	S/D
4	Emaya	S/D	S/D	S/D	S/D	S/D	S/D	S/D	1	S/D
...	...	...	...	...	...	...	...	...	...	...
123	Tecnico I.P.C. Andres Rosa	1	S/D	S/D	S/D	S/D	S/D	1	6	S/D
124	TECNICO SUPERIOR	S/D	S/D	S/D	S/D	S/D	S/D	10	5	1
125	Los Martires	S/D	S/D	S/D	S/D	S/D	S/D	S/D	S/D	S/D
126	Tecnico Superior	S/D	S/D	S/D	S/D	S/D	S/D	10	5	1
127	Total general	8	14	6	6	2	14	275	481	20

128 rows × 22 columns

Fuente: https://dev.socrata.com/foundry/www.datos.gov.co/52mm-fccv

JOSE FERNANDO GALINDO SUAREZ

G R A C I A S

Línea de atención al ciudadano: 01 8000 910270
Línea de atención al empresario: 01 8000 910682

@SENAComunica

www.sena.edu.co