Manipulación y visualización de datos (básicas)

A lo largo de este Jupyter Notebook explicaré algunos básicos sobre manipulación y visualización de datos con Python. Los tópicos específicos a cubrir en este cuaderno interactivo son los siguientes:

- Formatos de datos
- Cargar una base de datos
- Extraer datos de una base
- Estadísticos descriptivos
- Visualizaciones básicas

Repaso rápido a formatos de datos...

Puedes consultar la siguiente referencia: https://5stardata.info/es/ (https://5stardata.info/es/)

Cargando una base de datos

Para cargar una base de datos, utilizaremos un paquete de Python que resulta muy útil y eficiente, llamado Pandas (https://pandas.pydata.org). Dado que a los datos podemos encontrarlos en la misma ruta que este cuaderno interactivo, bastará sólo usar la función `.read_excel()` de Pandas.

In [1]:
```
import pandas

nombre_de_archivo = 'puntosconectividadleonjun2019.xlsx'
dataframe = pandas.read_excel(nombre_de_archivo)
```

Podemos imprimir toda la base de datos llamando a la variable `dataframe` , o acceder sólo al inicio (o final) de los datos con `.head()` (o `.tail()`).

In [2]:
```
dataframe.head()
```

Out[2]:

	ID	PUNTO_DE_CONEXION	DOMICILIO	UBICACION	LONGITUD	LATITUD	TIPO	ZONA
0	1	Primaria 5 de mayo	Boulevard Renacimiento del Potrero	Nuevo Amanecer	-101.607069	21.117987	colonia	urbana
1	2	Primaria Joel Cisneros Lara	Boulevard Torre Leon SN	Paseo de las Torres	-101.758375	21.167274	colonia	urbana
2	3	Primaria Tierra y Libertad	Ejido Los Naranjos Oficinas del Partido Accion...	Los Naranjos	-101.618882	21.154513	colonia	urbana
3	4	Primaria Dr Pablo del Rio	Boulevard Campestre Ahuehuete SN	Urbi Villa del Roble	-101.765204	21.166462	colonia	urbana
4	5	Primaria Miguel Hidalgo	Valle del Sahuan SN	Rancho Nuevo El Maguey	-101.620351	21.085333	colonia	urbana

Tanto `.head()` como `.head()` de manera estándar retorna 5 elementos de la tabla de datos, sin embargo pueden recibir como argumento un número entero para devolver esa cantidad de elementos.

Ejercicio:

Haz que las funciones ya mecnionadas devuelvan una cantidad distinta de elementos.

In [3]:
```
# TODO.
# Haz que del dataframe .head() devuelva los primeros 10 elementos

dataframe.head(10)
```

Out[3]:

	ID	PUNTO_DE_CONEXION	DOMICILIO	UBICACION	LONGITUD	LATITUD	TIPO	ZC
0	1	Primaria 5 de mayo	Boulevard Renacimiento del Potrero	Nuevo Amanecer	-101.607069	21.117987	colonia	urb
1	2	Primaria Joel Cisneros Lara	Boulevard Torre Leon SN	Paseo de las Torres	-101.758375	21.167274	colonia	urb
2	3	Primaria Tierra y Libertad	Ejido Los Naranjos Oficinas del Partido Accion...	Los Naranjos	-101.618882	21.154513	colonia	urb
3	4	Primaria Dr Pablo del Rio	Boulevard Campestre Ahuehuete SN	Urbi Villa del Roble	-101.765204	21.166462	colonia	urb
4	5	Primaria Miguel Hidalgo	Valle del Sahuan SN	Rancho Nuevo El Maguey	-101.620351	21.085333	colonia	urb
5	6	Telesecundaria num 581	Antigua Salida A San Felipe No 103	Ibarrilla	-101.652170	21.187295	colonia	urb
6	7	Primaria Emiliano Zapata	Emiliano Zapata km 14	Los Sauces	-101.541238	21.024297	comunidad	r
7	8	Telesecundaria No 123	Estudiante No 1 Carretera Leon Silao	Los Sauces	-101.677634	21.170842	comunidad	r
8	9	Primaria Albino Garcia	Comunidad San Jose del Potrero Club Cinegetico...	San Jose del Potrero	-101.600516	21.134267	comunidad	r
9	10	Primaria Insurgentes	Canada de Alfaro pasando el rio	Alfaro	-101.608377	21.148399	comunidad	r

In [4]:
```
# TODO.
# Ahora haz que del dataframe .tail() devuelva los ultimos 2 elementos

dataframe.tail(2)
```

Out[4]:

	ID	PUNTO_DE_CONEXION	DOMICILIO	UBICACION	LONGITUD	LATITUD	TIPO	Z
64	65	Santa Rosa Plan de Ayala	En el Templo y Comandancia de policia	Santa Rosa Plan de Ayala	-101.722211	21.070953	comunidad	
65	66	El Terrero	Calle Alamo 3	El Terrero	-101.605278	20.961111	comunidad	

Extraer datos de una base

Para extraer las cabeceras de una tabla de datos, podemos convertir a lista una tabla misma:

In [5]:
```
cabeceras = list(dataframe)
cabeceras
```

Out[5]:
```
['ID',
 'PUNTO_DE_CONEXION',
 'DOMICILIO',
 'UBICACION',
 'LONGITUD',
 'LATITUD',
 'TIPO',
 'ZONA',
 'LOCALIDAD',
 'SEDE',
 'FECHA_INSTAL',
 'POBLACION']
```

In [6]:
```
print(cabeceras[1])
```

```
PUNTO_DE_CONEXION
```

Si quisiéramos extraer una columna completa de una tabla de datos, podemos utilizar la cabecera como sigue:

In [7]:
```
poblacion = dataframe[[ 'POBLACION' ]]
poblacion
```

Out[7]:

	POBLACION
0	486
1	247
2	201
3	689
4	447
...	...
61	2905
62	1698
63	857
64	5134
65	208

66 rows × 1 columns

De aquí podemos observar que la extracción de objetos conserva la estructura de la misma tabla de datos, por lo que para transofrmar los valores a formato de lista, utilizamos el atributo `.values` .

Una vez transformado en lista podemos utilizar indexación como vimos previamente.

In [8]:
```
poblacion.values
```

Out[8]:
```
array([[ 486],
       [ 247],
       [ 201],
       [ 689],
       [ 447],
       [ 318],
       [ 384],
       [ 323],
       [ 611],
       [ 410],
       [  46],
       [ 230],
       [ 238],
       [ 292],
       [ 136],
       [   0],
       [   0],
       [   0],
       [2311],
       [2381],
       [  54],
       [ 147],
       [ 137],
       [ 141],
       [  19],
       [ 863],
       [  73],
       [  47],
       [ 137],
       [ 101],
       [  47],
       [ 974],
       [ 810],
       [  86],
       [6261],
       [  77],
       [1249],
       [ 280],
       [ 880],
       [ 288],
       [ 152],
       [ 513],
       [ 448],
       [1567],
       [ 788],
       [ 797],
       [2136],
       [1642],
       [ 552],
       [1174],
       [ 652],
       [2473],
       [1228],
       [2875],
       [1218],
       [ 896],
       [ 827],
```

```
[ 417],
[1025],
[   0],
[ 160],
[2905],
[1698],
[ 857],
[5134],
[ 208]])
```

Puedes filtrar datos utilizando condicionales dentro de los corchetes de selección.

Ejemplo:

Si quisiéramos filtrar a los puntos de conectividad que benefician a una población de más de 2000 personas, hacemos:

In [9]:
```
dataframe[dataframe['POBLACION'] > 2000]
```

Out[9]:

	ID	PUNTO_DE_CONEXION	DOMICILIO	UBICACION	LONGITUD	LATITUD	TIPO	Z
18	19	San Francisco de Duran	Alvaro Obregon	San Agustin de Miraflores	-101.623439	21.004776	comunidad	
19	20	Alfaro	Templo de Alfaro	Alfaro	-101.612177	21.147961	comunidad	
34	35	Duarte	Calle La luz SN	Parroquia del Senor de la Misericordia	-101.522203	21.085650	comunidad	
46	47	La Sandia	En el Templo	La Sandia	-101.697000	20.922353	comunidad	
51	52	Los Ramirez	Calle principal SN	Los Ramirez	-101.645933	21.019667	comunidad	
53	54	Loza de los Padres	Calle principal SN	Loza de los Padres	-101.547222	21.071667	comunidad	
61	62	San Juan de Otates	Calle principal SN	San Juan de Otates	-101.557783	21.114464	comunidad	
64	65	Santa Rosa Plan de Ayala	En el Templo y Comandancia de policia	Santa Rosa Plan de Ayala	-101.722211	21.070953	comunidad	

Ejercicio:

Filtra elementos de acuerdo a la zona, específicamente si es `'rural'`.

In [10]:
```
# TODO.
# Filtra elementos con la condición 'rural' que pertenezcan a la columna
'ZONA'

dataframe[dataframe['ZONA'] == 'rural']
```

Out[10]:

	ID	PUNTO_DE_CONEXION	DOMICILIO	UBICACION	LONGITUD	LATITUD	TIPO
6	7	Primaria Emiliano Zapata	Emiliano Zapata km 14	Los Sauces	-101.541238	21.024297	comunidad
7	8	Telesecundaria No 123	Estudiante No 1 Carretera Leon Silao	Los Sauces	-101.677634	21.170842	comunidad
8	9	Primaria Albino Garcia	Comunidad San Jose del Potrero Club Cinegetico...	San Jose del Potrero	-101.600516	21.134267	comunidad
9	10	Primaria Insurgentes	Canada de Alfaro pasando el rio	Alfaro	-101.608377	21.148399	comunidad
10	11	Telesecundaria No 1003	Mesa de Ibarrilla Salida a Ibarrilla SN	Mesa de Ibarrilla	-101.650096	21.219478	comunidad
11	12	Primaria Benito Juarez	Camino viejo a Lagos SN y griega carretera Leo...	Lagunillas	-101.774393	21.186584	comunidad
12	13	Telesecundaria Num 527	Camino a los Tepetates No 4	Los Arcos	-101.674966	21.065954	comunidad
13	14	Telesecundaria num 528	Camino Calle de los volcanes No 102	San Pedro del Monte Hospital	-101.711722	21.033523	comunidad
14	15	Primaria Melchor Ocampo	Carretera Leon San Francisco Del Rincon	La Mora	-101.763421	21.065627	comunidad
18	19	San Francisco de Duran	Alvaro Obregon	San Agustin de Miraflores	-101.623439	21.004776	comunidad
19	20	Alfaro	Templo de Alfaro	Alfaro	-101.612177	21.147961	comunidad
20	21	Los Alisos	Casa del guardabosques y comedor comunitario d...	Los Alisos	-101.648564	21.134395	comunidad
21	22	Las Canelas	Iglesia de la comunidad	Las Canelas	-101.466532	21.218908	comunidad
22	23	Cuesta Blanca	Terreno ubicado a un costado de la calle princ...	Cuesta Blanca	-101.491160	21.105861	comunidad
23	24	El Derramadero	A un costado de la Escuela Primaria Francisco ...	El Derramadero	-101.416859	21.167080	comunidad
24	25	La Mesa del Obispo	Frente a comedor comunitario	Mesa del Obispo	-101.435120	21.147660	comunidad
25	26	Nuevo Valle de Moreno	Plaza Principal	Nuevo Valle de Moreno	-101.425030	21.210768	comunidad

	ID	PUNTO_DE_CONEXION	DOMICILIO	UBICACION	LONGITUD	LATITUD	TIPO
26	27	San Jose de Otates Sur	Comedor comunitario de la comunidad	San Jose de Otates Sur	-101.578753	21.260250	comunidad
27	28	San Jose de Otates Norte	Escalinata para cceder al atrio de la comunidad	San Jose de Otates Norte	-101.804239	20.781932	comunidad
28	29	San Jose de los Romeros	Calle Juarez 60	San Jose de los Romeros	-101.492481	21.037170	comunidad
29	30	Sauz Seco	Terreno frente al atrio de la iglesia	Sauz Seco	-101.545562	21.168352	comunidad
30	31	San Rafael Cerro Verde	Plaza Principal	San Rafael Cerro Verde	-101.460000	21.162740	comunidad
31	32	Vaquerias	En la Plaza principal frente a la iglesia	Vaquerias	-101.401757	21.155356	comunidad
32	33	Albarradones	Calle Central SN	Parroquia de Nuestra SeÒora del Refugio	-101.503611	21.040278	comunidad
33	34	Benito de Juarez	Juan de Grijalva 116	Benito de Juarez	-101.580278	21.077500	comunidad
34	35	Duarte	Calle La luz SN	Parroquia del Senor de la Misericordia	-101.522203	21.085650	comunidad
35	36	Ejido Pompa	Calle Lazaro Cardenas 202	Ejido Pompa	-101.687500	21.075278	comunidad
36	37	El Capricho	Plaza de la comunidad	El Capricho	-101.690000	21.997778	comunidad
37	38	El Nacimiento	Calle Privada Santa Maria 128	El Nacimiento	-101.772222	21.058333	comunidad
38	39	El Ramillete	Calle Flor de melocoton SN	El Ramillete	-101.591462	21.043085	comunidad
39	40	El Vergel	Calle Vergel SN	El Vergel	-101.690000	21.075000	comunidad
40	41	Granjas Economicas	Avenida del sur 60	Lomas del suspiro	-101.553611	21.034722	comunidad
41	42	Guadalupe Victoria	Calle Principal 213	Guadalupe Victoria	-101.579722	21.019444	comunidad
42	43	Jacales	Calle Alamo SN	Jacales	-101.548889	21.036944	comunidad
43	44	La Laborcita	Calle principal 113	Primaria Justo Sierra	-101.557662	21.106974	comunidad
44	45	La Patina	Calle principal al Templo SN	La Patina	-101.705556	21.195556	comunidad

	ID	PUNTO_DE_CONEXION	DOMICILIO	UBICACION	LONGITUD	LATITUD	TIPO
45	46	La Providencia	Calle primero de enero 6	La Providencia	-101.660000	21.074167	comunidad
46	47	La Sandia	En el Templo	La Sandia	-101.697000	20.922353	comunidad
47	48	Ladrillera del Refugio	Plaza del templo	Ladrillera del Refugio	-101.553611	21.084722	comunidad
48	49	Lagunillas	Calle Andador del Caserio 104	Lagunillas	-101.765000	21.202222	comunidad
49	50	Los Arcos	Calle principal 25	Los Arcos	-101.687222	21.049167	comunidad
50	51	Los Lopez	Calle principal 255	Los Lopez	-101.570314	21.047886	comunidad
51	52	Los Ramirez	Calle principal SN	Los Ramirez	-101.645933	21.019667	comunidad
52	53	Los Sauces	Calle Lopez Mateos SN	Los Sauces	-101.539381	21.023478	comunidad
53	54	Loza de los Padres	Calle principal SN	Loza de los Padres	-101.547222	21.071667	comunidad
54	55	Lucio Blanco	Lucio Blanco SN Cuenca Rio lerma	Lucio Blanco	-101.551111	21.115278	comunidad
55	56	Nuevo Lindero	Calle Ignacio Medina 20	Nuevo Lindero	-101.636944	21.958889	comunidad
56	57	Puerta de San German	Calle Feliciano Domiguez	Puerta de San German	-101.780278	21.033611	comunidad
57	58	Puerta del Cerro	Calle Carlos Medina SN	Puerta del Cerro	-101.786111	21.050000	comunidad
58	59	Rancho Nuevo La Luz	En el Templo	Rancho Nuevo La Luz	-101.645000	21.963056	comunidad
59	60	San Carlos La Roncha	Casa particular a un lado del templo	San Carlos La Roncha	-101.587267	21.071008	comunidad
60	61	San JoseÇ del Barron El Cachete	Calle principal SN	San Jose del Barron El Cachete	-101.496389	21.161944	comunidad
61	62	San Juan de Otates	Calle principal SN	San Juan de Otates	-101.557783	21.114464	comunidad
62	63	San Judas	Calle San Judas SN	Frente al Templo	-101.704722	21.999167	comunidad
63	64	San Pedro del Monte	Calle de las haciendas	San Pedro del Monte	-101.714722	21.031389	comunidad
64	65	Santa Rosa Plan de Ayala	En el Templo y Comandancia de policia	Santa Rosa Plan de Ayala	-101.722211	21.070953	comunidad
65	66	El Terrero	Calle Alamo 3	El Terrero	-101.605278	20.961111	comunidad

Estadísticos descriptivos

De manera muy general, Pandas ya puede entregarnos un resumen estadístico descriptivo utilizando funciones como `.count()` o `.describe()`.

Dichas descripciones contienen elementos estadísticos como la media, mímimos, máximos, percentiles (https://es.wikipedia.org/wiki/Percentil) y desviación estándar (https://es.wikipedia.org/wiki/Desviación_típica).

In [11]:
```
dataframe.count()
```

Out[11]:
```
ID                  66
PUNTO_DE_CONEXION   66
DOMICILIO           66
UBICACION           66
LONGITUD            66
LATITUD             66
TIPO                66
ZONA                66
LOCALIDAD           66
SEDE                66
FECHA_INSTAL        66
POBLACION           66
dtype: int64
```

In [12]:
```
dataframe.describe()
```

Out[12]:

	ID	LONGITUD	LATITUD	POBLACION
count	66.000000	66.000000	66.000000	66.000000
mean	33.500000	-101.621552	21.149655	829.742424
std	19.196354	0.101834	0.226045	1136.706136
min	1.000000	-101.804239	20.781932	0.000000
25%	17.250000	-101.695250	21.048206	142.500000
50%	33.500000	-101.619617	21.099640	432.000000
75%	49.750000	-101.551736	21.165531	954.500000
max	66.000000	-101.401757	21.999167	6261.000000

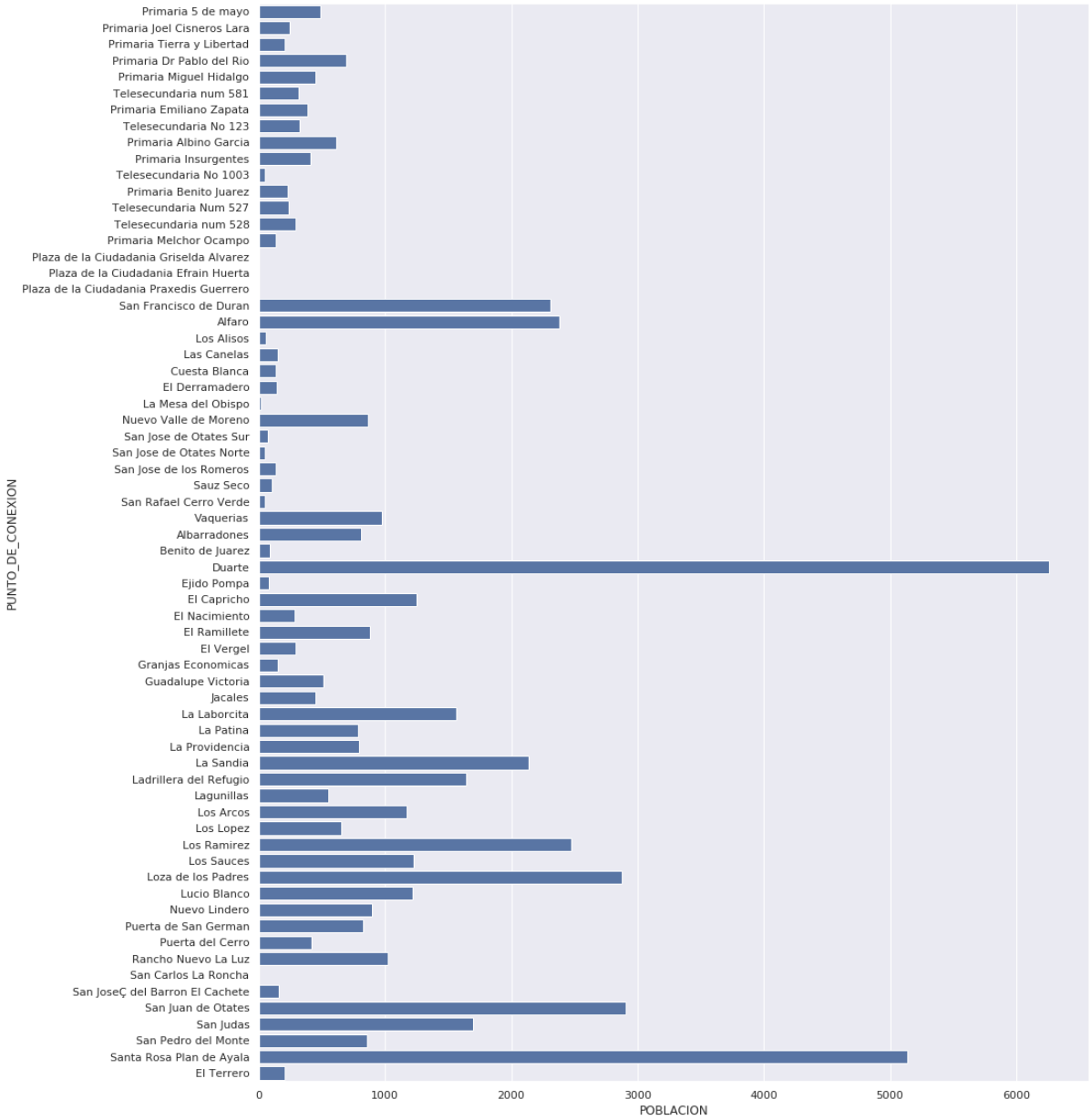
Visualizaciones básicas

Otro paquete que resulta muy útil, para gráficos estadísticos descriptivos, es Seaborn (https://seaborn.pydata.org).

In [13]:
```
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
sns.set(style="darkgrid")
```

In [17]:
```
plt.figure(figsize=(15, 20))
sns.barplot(x="POBLACION", y="PUNTO_DE_CONEXION", data=dataframe,
            label="Población beneficiada", color="b")
```

Out[17]: <matplotlib.axes._subplots.AxesSubplot at 0x7f4b7192e5f8>



Las visualizaciones y su utilidad dependen un poco de los datos, en este caso, dado que tenemos geo-referenciación, valdría la pena explorar los datos utilizando mapas. 🗺️ 👀

Sin embargo, te comparto el siguiente material para que puedas conocer más visualizaciones que resultan útiles:

- La galería de gráficos de Seaborn (https://seaborn.pydata.org/examples/index.html)
- La galería de gráficos en Python (https://python-graph-gallery.com)