

TIPOLOGÍA Y CICLO DE VIDA DE LOS DATOS – PRÁCTICA 1

Carlos Giner Baixauli

1. Contexto. Explicar en qué contexto se ha recolectado la información. Explique por qué el sitio web elegido proporciona dicha información.

La información ha sido recolectada de Wikipedia, que, según su propia definición, es una enciclopedia libre, políglota y editada de manera colectiva. Esta enciclopedia contiene información actualizada y precisa en el ámbito de la demografía, ya que se basa en fuentes fiables como los World Urbanization Prospects de Naciones Unidas, entre otras. La información recolectada consiste en una tabla con las ciudades más pobladas del mundo. La elección de este sitio web se debe a que integra en una única tabla la información deseada, de esta manera no será necesario hacer web scraping sobre múltiples fuentes.

2. Definir un título para el dataset. Elegir un título que sea descriptivo.

En dataset seleccionado se titula “Las ciudades más pobladas del mundo”, ya que recopila la información demográfica sobre las ciudades que tienen más de tres millones de habitantes.

3. Descripción del dataset. Desarrollar una descripción breve del conjunto de datos que se ha extraído (es necesario que esta descripción tenga sentido con el título elegido).

El dataset obtenido contiene 4 variables y 86 observaciones. Las variables son el orden en el ranking, el nombre de la ciudad, el número de habitantes de la ciudad, y el país en el que se encuentra. La información más interesante que se puede obtener del dataset es el número de habitantes, ya que este será el criterio principal para comparar las ciudades entre sí.

4. Representación gráfica. Presentar una imagen o esquema que identifique el dataset visualmente.

A continuación, mostramos un gráfico de barras para visualizar la información contenida en el dataset. La longitud de las barras indica la población de las ciudades, y el color será distinto según el país al que pertenecen.

7. Inspiración. Explique por qué es interesante este conjunto de datos y qué preguntas se pretenden responder.

El interés de este dataset consiste en la posibilidad de comparar las grandes ciudades del mundo entre sí en términos de población, así como de agrupar dichas ciudades por países. Puede ser de gran utilidad para la realización de estudios demográficos tanto a nivel regional como global, así como para identificar las regiones más pobladas para la toma de decisiones en otros ámbitos como el de la economía.

8. Licencia. Seleccione una de estas licencias para su dataset y explique el motivo de su selección:

Released Under CC0: Public Domain License

Hemos seleccionado esta licencia para facilitar el uso libre puesto que ofrecemos el dataset al dominio público, liberándolo de todos los derechos de propiedad intelectual, incluyendo todos los derechos conexos. Pueden copiar, modificar, distribuir la obra y hacer comunicación pública, incluso para fines comerciales.

9. Código. Adjuntar el código con el que se ha generado el dataset, preferiblemente en Python o, alternativamente, en R.

```
# Importamos las librerías
import pandas as pd
import requests
from bs4 import BeautifulSoup
import re

# Obtenemos el texto en xml de la página web
web =
requests.get('https://en.wikipedia.org/wiki/List_of_cities_proper_by_p
opulation').text
soup = BeautifulSoup(web, 'xml')
print(soup.prettify())

# Buscamos la tabla y sus filas en el documento
table = soup.find('table', {'class':'sortable wikipable mw-datatable'})
table_rows = table.find_all('tr')

# Guardamos la información en un dataset
data = []
for row in table_rows:
    data.append([t.text.strip() for t in row.find_all('td')])

# Obtenemos los nombres de las columnas
columns = []
columns.append([t.text.strip() for t in table_rows[0].find_all('th')])
columns = columns[0]
```

```
# Filtramos las columnas que nos interesan
df = pd.DataFrame(data, columns=columns)
df = df[['City', 'Population', 'Country']]

# Limpiamos el dataset
df = df.drop(df.index[0])
df['Population'] = df['Population'].str.replace(r'\[.*?\]', '')
df['Population'] = df['Population'].str.replace(r',', '')
df['Population'] = pd.to_numeric(df['Population'])

#Guardamos el dataset en formato csv
df.to_csv("dataset.csv")
```

10. Dataset. Presentar el dataset en formato CSV

```
,City,Population,Country
1,Chongqing,30165500,China
2,Shanghai,24183300,China
3,Beijing,21707000,China
4,Istanbul,15029231,Turkey
5,Karachi,14910352,Pakistan
6,Dhaka,14399000,Bangladesh
7,Tokyo,13515271,Japan
8,Moscow,13200000,Russia
9,Guangzhou,13081000,China
10,Shenzhen,12528300,China
11,Mumbai,12442373,India
12,São Paulo,12252023,Brazil
13,Kinshasa,11462000,Democratic Republic of the Congo
14,Tianjin,11249000,China
15,Lahore,11126000,Pakistan
16,Delhi,6787941,India
17,Jakarta,10624000,Indonesia
18,Dongguan,10615000,China
19,Seoul,9806000,South Korea
20,Foshan,9279000,China
21,Chengdu,9012000,China
22,Lima,8894000,Peru
```

23,Mexico City,8918653,Mexico
24,Tehran,9033003,Iran
25,London,8825001,United Kingdom
26,Bangalore,8443675,India
27,New York,8398748,United States
28,Shenyang,8106171,China
29,Wuhan,8035000,China
30,Bogotá,7963000,Colombia
31,Ningbo,7605689,China
32,Cairo,7601018,Egypt
33,Ho Chi Minh City,7431000,Vietnam
34,Hong Kong,7298600,China
35,Nanjing,7260000,China
36,Hanoi,7232700,Vietnam
37,Changsha,7044118,China
38,Hangzhou,7035000,China
39,Hyderabad,6993262,India
40,Baghdad,6793000,Iraq
41,Chennai,6727000,India
42,Riyadh,6694000,Saudi Arabia
43,Rio de Janeiro,6520000,Brazil
44,Xi'an,6220000,China
45,Suzhou,5983000,China
46,Bangkok,5782000,Thailand
47,Singapore,5607000,Singapore
48,Ahmedabad,5570585,India
49,Santiago Metropolitan,5561000,Chile
50,Shantou,5391028,China
51,Saint Petersburg,5351000,Russia
52,Harbin,5299000,China
53,Dar es Salaam,5257000,Tanzania
54,Yangon,5214000,Myanmar
55,Johannesburg,4949000,South Africa
56,Abidjan,4765000,Ivory Coast
57,Alexandria,4616625,Egypt

58, Kolkata, 4496694, India
 59, Ankara, 4470800, Turkey
 60, Surat, 4467797, India
 61, Giza, 4239988, Egypt
 62, Zhengzhou, 4122087, China
 63, Los Angeles, 3976322, United States
 64, New Taipei City, 3954929, Taiwan
 65, Cape Town, 3740026, South Africa
 66, Yokohama, 3726167, Japan
 67, Berlin, 3671000, Germany
 68, Busan, 3590000, South Korea
 69, Xiamen, 3531347, China
 70, Jeddah, 3456259, Saudi Arabia
 71, Durban, 3442361, South Africa
 72, Kabul, 3414100, Afghanistan
 73, Casablanca, 3359818, Morocco
 74, Hefei, 3352076, China
 75, Pyongyang, 3255388, North Korea
 76, Madrid, 3207247, Spain
 77, Faisalabad, 3203846, Pakistan
 78, Ekurhuleni, 3178470, South Africa
 79, Nairobi, 3138369, Kenya
 80, Zhongshan, 3121275, China
 81, Pune, 3115431, India
 82, Addis Ababa, 3103673, Ethiopia
 83, Jaipur, 3073350, India
 84, Buenos Aires, 3054300, Argentina
 85, Wenzhou, 3039439, China
 86, Incheon, 3002645, South Korea

Contribuciones	Firma
Investigación previa	CGB
Redacción de las respuestas	CGB
Desarrollo código	CGB