M2.851 - Tipología y ciclo de vida de los datos
PRAC 1
Diego Cheuquepán Maldonado

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

   Estamos en el ámbito del mercado ganadero, en específico en la producción de carne bovina que en Chile que se ubica en el tercer lugar de producción, tras las carnes de ave y de cerdo (https://www.odepa.gob.cl/rubros/carnes).

   En este contexto, la producción de carne de vacuno en Chile se orienta principalmente al mercado interno y es un mercado muy relevante ya que a diferencia de otros como el avícola o porcino, es un mercado no concentrado, contando con más de 120 mil productores.

   Dichos productores comercializan el vacuno a través de ferias ganaderas que están distribuidas a lo largo del país y que semana a semana reportan los precios de venta de las diferentes categorías.

   A partir de esto, el proyecto consiste en obtener la información del precio de venta que informa semanalmente la feria de ganaderos de Osorno S.A. (FEGOSA) con el propósito de obtener los datos de comercialización de las diferentes categorías ganaderas que tiene dicho consorcio (bovino, caprino, ovino, equino, etc).

   Para ello, se utilizará la técnica de web scraping para obtener dicha información desde el portal de FEGOSA (http://www.fegosa.cl/precios.html) que semana a semana publica sus precios de venta de las ferias Osorno, Paillaco, Puerto Montt y Puerto Varas, que un día a la semana se lleva a cabo para comercializar el ganado.

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

   La revisión del boletín de carne bovina de la Oficina de Estudios y Políticas Agrarias (ODEPA) nos entrega información agregada sobre diferentes indicadores del mercado bovino.

   A partir de dicho análisis, la propuesta de título del dataset es:

   "Precios de ganado reportados por ferias FEGOSA"

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).**

Considerando los diferentes atributos del dataset y que estos no fueron limpiados ni tratados, a continuación, presento los principales datos descriptivos.

Atributo feria

- Tipo de dato: Cualitativo - Nominal
- N: 4

Atributo fecha

- Tipo de dato: Cualitativo – Nominal
- N: 16

Atributo especie

- Tipo de dato: Cualitativo – Nominal
- N: 20

Atributo cabezas

- Tipo de dato: Cuantitativo – Discreto
- N: 11.868

Atributo peso promedio

- Tipo de dato: Cuantitativo – Continuo
- N: 276

Atributo promedio 5 primeros precios

- Tipo de dato: Cuantitativo – Continuo
- N: 276

Atributo promedio general

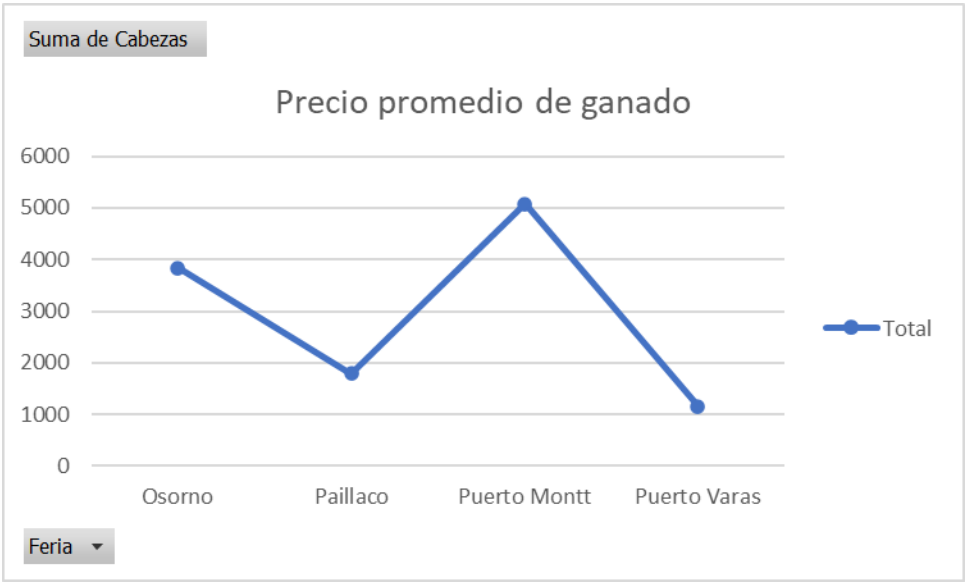
- Tipo de dato: Cuantitativo – Continuo
- N: 276

Feria	N Cabezas	Peso promedio	Precio promedio
Osorno	3.835	405,67	$782,04
Paillaco	1.793	388,72	$656,52
Puerto Montt	5.077	417,50	$824,94
Puerto Varas	1.163	364,03	$820,44

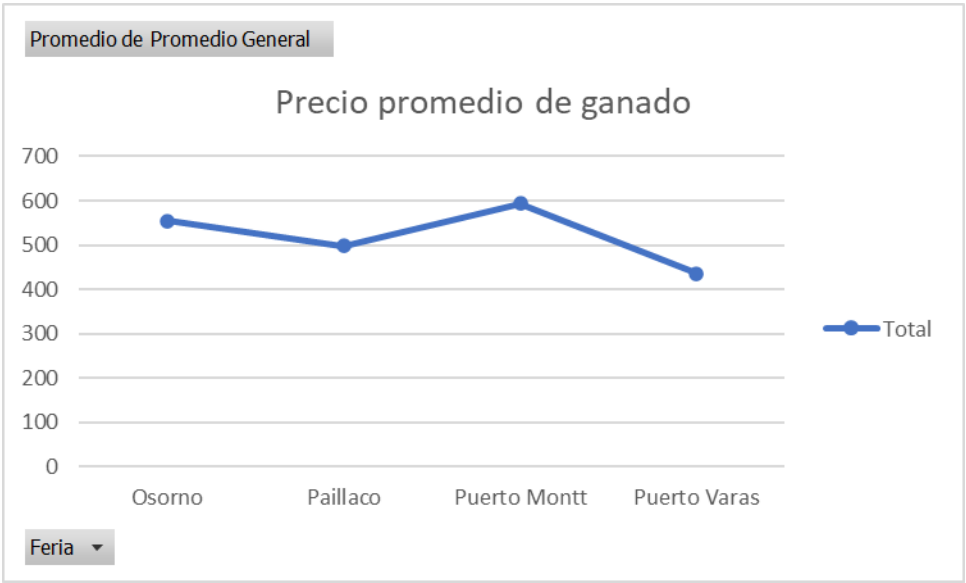
4. **Representación gráfica. Presentar una imagen o esquema que identifique el dataset visualmente.**

El dataset contiene datos respecto a las cabezas de ganado vendidas por las ferias de FEGOSA.

El siguiente gráfico muestra las cabezas de ganado vendido por las ferias Osorno, Paillaco, Puerto Montt y Puerto Varas.



En tanto, el siguiente gráfico muestra el precio promedio por kilo de ganado vendido en las ferias.



Por último, podemos obtener la especie de ganado por feria.

5. **Contenido. Explicar los campos que incluye el dataset, el periodo de tiempo de los datos y cómo se ha recogido.**

El dataset contiene la información del precio de venta promedio de las diferentes categorías de ganado que se comercializan en las ferias FEGOSA (feria de ganaderos de Osorno), principal consorcio de comercialización ovina de la Región de Los Lagos y Aysén.

Dicha agrupación cuenta con 4 ferias que se llevan a cabo durante distintos días de la semana, siguiendo el siguiente calendario:

LUNES	MARTES	MIÉRCOLES	JUEVES	VIERNES
Osorno	Paillaco	Puerto Montt	*Sin feria*	Puerto Varas

El dataset contiene:

- Feria: Nombre de la feria de FEGOSA donde se realizó la venta de ganado un determinado día.
- Fecha: Día en que se lleva a cabo la feria y en la cual se comercializa el ganado.
- Especie: Categoría de ganado, según los criterios de ODEPA Chile.
- N° de cabezas: Cantidad de animales de cada categoría que se comercializan el día de la feria.
- Peso promedio: Media aritmética de los pesos de cada cabeza de ganado de una determinada categoría.
- Precio 1: Corresponde al primer precio de venta.
- Precio 2: Corresponde al segundo precio de venta.
- Precio 3: Corresponde al tercer precio de venta.
- Precio 4: Corresponde al cuarto precio de venta.
- Precio 5: Corresponde al quinto precio de venta.
- Promedio 5 PP: Corresponde al precio promedio de los primeros 5 precios de venta.
- Promedio general: Corresponde al precio promedio de venta.

Respecto al proceso de extracción se comenzó con la revisión del sitio web de FEGOSA (http://www.fegosa.cl/index.html).
Además, se realizó una revisión de la tecnología, el dueño del sitio y el archivo robots.

Sobre la tecnología, el sitio web está construido con DreamWeaver y está en un servidor web Apache.

Código

```
15  # Identificamos la tecnología del sitio web
16  import builtwith
17  print("")
18  print("Información de la tecnología del sitio web FEGOSA")
19  print(builtwith.parse(url))
20  print("================================================")
```

Resultado

```
Información de la tecnología del sitio web FEGOSA
{'web-servers': ['Apache'], 'editors': ['DreamWeaver']}
================================================
```

Sobre el dueño del sitio confirmamos que se trata de FEGOSA SA.

Código

```
22  # Identificamos al dueño del sitio web
23  import whois
24  print("")
25  print("Información del dueño del sitio web FEGOSA")
26  print(whois.whois(url))
27  print("================================================")
```

Resultado

```
================================================

Información del dueño del sitio web FEGOSA
{
  "domain_name": "fegosa.cl",
  "registrant_name": "Feria Ganaderos Osorno S.A. (FERIA GANADEROS OSORNO S A)",
  "registrant_organization": null,
  "registrar": "NIC Chile",
  "registrar_url": "https://www.nic.cl",
  "creation_date": "2000-05-03 22:09:22",
  "expiration_date": "2020-05-30 19:09:22",
  "name_servers": [
    "ns.gtdinternet.com",
    "ns2.gtdinternet.com"
  ]
}
================================================
```

Finalmente, al buscar el archivo robots.txt no se encontraron resultados.

Código

```
# Análisis de robot.txt
import requests
response = requests.get("http://www.fegosa.cl/robots.txt")
test = response.text
print("")
print(response.status_code)
print("robots.txt del sitio web FEGOSA")
print(test)
print("==================================================")

```

Resultado

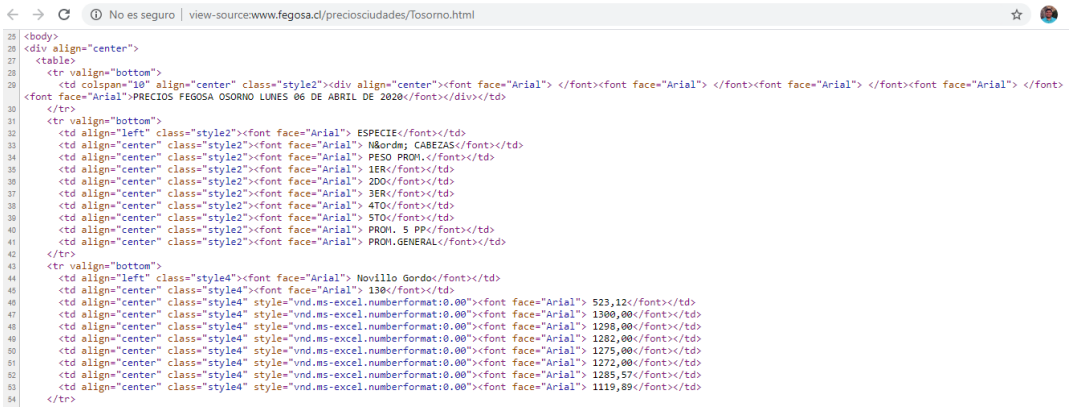
```
==================================================

404
robots.txt del sitio web FEGOSA
<!DOCTYPE HTML PUBLIC "-//IETF//DTD HTML 2.0//EN">
<html><head>
<title>404 Not Found</title>
</head><body>
<h1>Not Found</h1>
<p>The requested URL /robots.txt was not found on this server.</p>
</body></html>

==================================================
```

Luego, se revisó el sitio web buscando hallar el lugar donde se informan los precios. Una vez encontrado esta página (http://www.fegosa.cl/precios.html), se procedió a analizar su estructura con la funcionalidad *ver código fuente de la página* del navegador Chrome.

La página web, contiene el detalle de la última feria realizada en cada una de las ferias de FEGOSA según el calendario indicado anteriormente.

El análisis nos permitió comprender que cada feria tiene asociado una url que corresponde a una página HTML con la tabla de precios de la fecha informada.

Luego, se revisó cada tabla de precios usando la funcionalidad *ver código fuente de la página* del navegador Chrome con el propósito de revisar la estructura de etiquetas HTML.

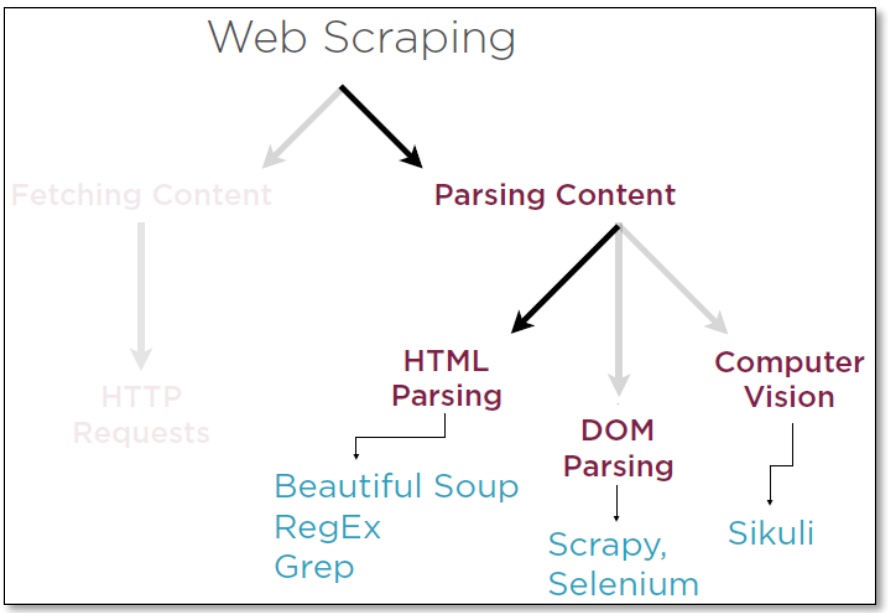



```
<body>
<div align="center">
  <table>
    <tr valign="bottom">
      <td colspan="10" align="center" class="style2"><div align="center"><font face="Arial"> </font><font face="Arial"> </font><font face="Arial"> </font><font face="Arial"> </font>
<font face="Arial">PRECIOS FEGOSA OSORNO LUNES 06 DE ABRIL DE 2020</font></div></td>
    </tr>
    <tr valign="bottom">
      <td align="left" class="style2"><font face="Arial"> ESPECIE</font></td>
      <td align="center" class="style2"><font face="Arial"> N&ordm; CABEZAS</font></td>
      <td align="center" class="style2"><font face="Arial"> PESO PROM.</font></td>
      <td align="center" class="style2"><font face="Arial"> 1ER</font></td>
      <td align="center" class="style2"><font face="Arial"> 2DO</font></td>
      <td align="center" class="style2"><font face="Arial"> 3ER</font></td>
      <td align="center" class="style2"><font face="Arial"> 4TO</font></td>
      <td align="center" class="style2"><font face="Arial"> 5TO</font></td>
      <td align="center" class="style2"><font face="Arial"> PROM. 5 PP</font></td>
      <td align="center" class="style2"><font face="Arial"> PROM.GENERAL</font></td>
    </tr>
    <tr valign="bottom">
      <td align="left" class="style4"><font face="Arial"> Novillo Gordo</font></td>
      <td align="center" class="style4"><font face="Arial"> 130</font></td>
      <td align="center" class="style4" style="vnd.ms-excel.numberformat:0.00"><font face="Arial"> 523,12</font></td>
      <td align="center" class="style4" style="vnd.ms-excel.numberformat:0.00"><font face="Arial"> 1300,00</font></td>
      <td align="center" class="style4" style="vnd.ms-excel.numberformat:0.00"><font face="Arial"> 1298,00</font></td>
      <td align="center" class="style4" style="vnd.ms-excel.numberformat:0.00"><font face="Arial"> 1282,00</font></td>
      <td align="center" class="style4" style="vnd.ms-excel.numberformat:0.00"><font face="Arial"> 1275,00</font></td>
      <td align="center" class="style4" style="vnd.ms-excel.numberformat:0.00"><font face="Arial"> 1272,00</font></td>
      <td align="center" class="style4" style="vnd.ms-excel.numberformat:0.00"><font face="Arial"> 1285,57</font></td>
      <td align="center" class="style4" style="vnd.ms-excel.numberformat:0.00"><font face="Arial"> 1119,89</font></td>
    </tr>
```

Esta revisión confirmó que:

i) Cada feria tiene un url propia.
ii) Cada url lleva a una página de precios.
iii) La página de precios es una tabla HTML.

Dado que las páginas solo eran HTML se utilizó la librería BeautifulSoup, ya que según indica Janani Ravi en Extracting Data from HTML with Beautiful Soup, esta herramienta es útil cuando se realiza parsing de contenido HTML.



Jananu Ravi (sf). Extracting Data from HTML with Beautiful Soup

Como el sitio de FEGOSA asocia una url a cada feria y en esta se informan los precios, se construyó una lista para almacenar las URL, de manera que luego pudiese recorrerse mediante un ciclo *for* y con ello conectarse a las diferentes páginas, optimizando el proceso de solicitud de datos.

```
62    newurl = [urlosorno, urlpaillaco, urlptomontt, urlptovaras]
63
64    for i in range(4):
65        x = newurl[i]
66        # Análisis the HTML
67        html = urlopen(x)
68        soup = BeautifulSoup(html.read(),"html5lib")
69        table = soup.find('table')
70        rows = table.findAll('tr')
71        divs = soup.findAll("table")
```

Luego, como cada página tiene una estructura similar, se creó un único algoritmo que captura cada dato de la tabla que almacena en una lista asociada a cada atributo.

Sin embargo, se tuvo la precaución de analizar cuando los datos eran de una u otra feria, para incorporar el nombre de la feria (Osorno, Paillaco, Puerto Montt o Puerto Varas). Para ello, se utilizó la sentencia selectiva *if* y se buscaron caracteres de cada nombre de feria para determinar un punto de comparación, según se muestra a continuación.

```
73        for div in divs:
74            rows = div.findAll('tr')
75            for row in rows[0:1]:
76                column = str(row.findAll('td')[0].text)
77                column = column.replace("PRECIOS FEGOSA ", "")
78                osorno = column.find("OSORNO")
79                paillaco = column.find("PAILLACO")
80                montt = column.find("NTT")
81                varas = column.find("VARAS")
82                if osorno == 4:
83                    fecha = column.replace("OSORNO", "")
84                if paillaco == 4:
85                    fecha = column.replace("PAILLACO", "")
86                if montt == 13:
87                    fecha = column.replace("PUERTO MONTT", "")
88                if varas == 11:
89                    fecha = column.replace("PUERTO VARAS", "")
```

Finalmente, se creó un data frame con los datos de las listas y este se volcó en un CSV que lleva por nombre fegosa.

```
179   import pandas as pd
180
181   # Definición archivo:
182   archivo = {'Feria': ciudad, 'Fecha': date, 'Especie': especie, 'Cabezas': cabezas,
183
184   # Creación DataFrame:
185   df_numeros = pd.DataFrame(archivo)
186
187   # Guarda datos en CSV:
188   df_numeros.to_csv('fegosa.csv', header=True, index=False)
189
```

6. **Agradecimientos. Presentar al propietario del conjunto de datos. Es necesario incluir citas de investigación o análisis anteriores (si los hay).**

La feria de ganaderos de Osorno (FEGOSA) es una empresa del sur de Chile, ubicada en las Regiones de Los Lagos y Aysén. Esta ofrece un servicio de remate de ganado bovino mediante sus ferias que operan entre lunes y viernes.

El análisis realizado en el marco de este proyecto consistió en revisar el portal de ODEPA Chile (Oficina de Estudios y Políticas Agrarias) con el propósito de conocer más sobre el marcado de carnes de Chile.

Luego de analizar los mercados de carne de ave, de cerdo, de vacuno y de otro ganado, se decidió analizar el mercado de ganado vacuno ya que es el de menor concentración en Chile.

El mercado de vacuno se lleva a cabo principalmente en ferias ganaderas a lo largo del país. Por ello, luego se revisó el portal de AFECH AG, asociación gremial de ferias ganaderas de Chile, con el propósito de conocer quiénes eran sus asociados y tener acceso a sus sitios web.
Los asociados hallados fueron los siguientes:

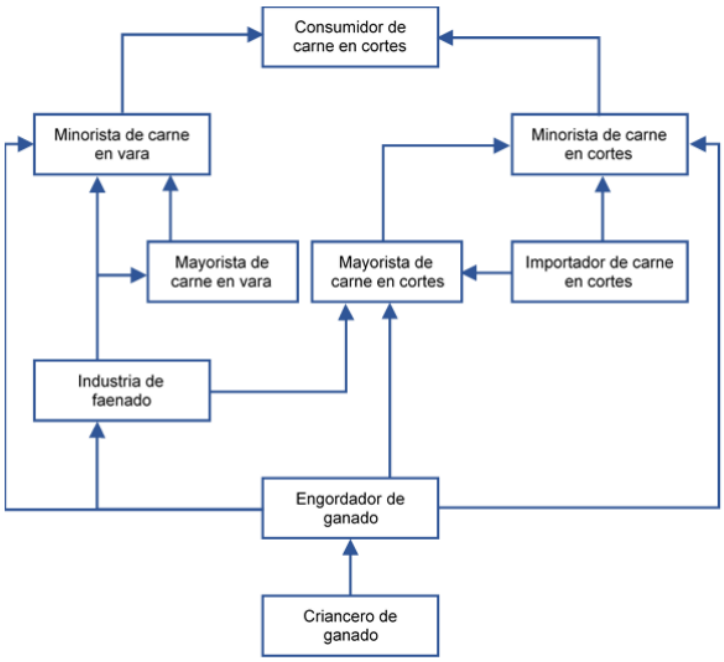
- Tattersal Ganado
- Fegosa
- Ferias Araucanía S.A.
- S.A. Feria de los Agricultores
- Ferias Bío-Bío Ltda.
- Feria Bernedo S.A.
- Feria Regional de Traiguén
- Feria Regional San Vicente Ltda.

No se encontraron análisis anteriores que se enfocaran en informar los datos desagregados de los precios de venta del ganado. Solo se encontraron memorias de las principales asociaciones ganaderas (FEGOSA, Tattersal) y de la misma ODEPA, pero en todos los casos se informaban los precios consolidados por año.

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

Mi familia es criancera de ganado y en la cadena de la carne bovina se encuentra en el eslabón más bajo según el diagrama propuesto por ODEPA (2016).

Al estar en el eslabón más bajo de toda la cadena productiva no cuentan con la información necesaria para asegurar un precio de venta que sea sustentable para la economía familiar y para evitar que los compradores de ganado para engorda (*Engordador de ganado*) utilicen su información y poder de compra que les permita acceder a precios infravalorados.



Fuente: https://www.odepa.gob.cl/wp-content/uploads/2017/12/cadenaCarneBovina.pdf
Página 7

En este escenario, las preguntas que busco responder mediante este proyecto es establecer la evolución del precio de venta del ganado por categoría teniendo información agregada de ciclos de tiempo, lo que permitiría establecer situaciones como las siguientes:

- Categoría con mejor/peor precio de venta.
- Meses del año con mejor/peor precio de venta.
- Tendencias al alza/baja de precios de venta.
- Predicción de precios futuros.

8. **Licencia. Seleccione una de estas licencias para su dataset y explique el motivo de su selección:**
    a. **Released Under CC0: Public Domain License**
    b. **Released Under CC BY-NC-SA 4.0 License**
    c. **Released Under CC BY-SA 4.0 License**
    d. **Database released under Open Database License, individual contents under Database Contents License**
    e. **Other (specified above)**
    f. **Unknown License**

    Entendiendo que la información relativa a los precios de venta es de acceso libre, se publican semanalmente a través del portal de la feria de ganaderos de Osorno (FEGOSA), y se consolidan en las memorias anuales de dicha agrupación ganadera, considero que la licencia a establecer para este dataset es la opción (b) "CC BY-NC-SA" ya que:

    i) Requiere el reconocimiento.
    ii) Permite la generación de obras derivadas.
    iii) No se puede utilizar la obra original con finalidades comerciales.
    iv) La distribución de las obras derivadas debe hacerse con una licencia igual a la de la obra original.

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

    Acceso al código: https://github.com/cheukepan/proyectoScraping

Código

```
"""
Created on Fri Apr 10 17:52:59 2020

@author: Diego Cheuquepán
"""

# Librerías
from urllib.request import urlopen
from bs4 import BeautifulSoup

# Definimos la URL del proyecto
url = 'http://www.fegosa.cl/preciososorno/precioosorno.html'

# Identificamos la tecnología del sitio web
import builtwith
print("")
print("Información de la tecnología del sitio web FEGOSA")
print(builtwith.parse(url))
print("====================================================")

# Identificamos al dueño del sitio web
import whois
print("")
print("Información del dueño del sitio web FEGOSA")
print(whois.whois(url))
print("====================================================")

# Análisis de robot.txt
import requests
```

```
response = requests.get("http://www.fegosa.cl/sitemap.xml")
test = response.text
print("")
print(response.status_code)
print("robots.txt del sitio web FEGOSA")
print(test)
print("==================================================")

# Definimos las matrices

especie = []
cabezas = []
pesoProm = []
primer = []
segundo = []
tercero = []
cuarto = []
quinto = []
promedio = []
general = []
date = []
ciudad = []

fecha = ""

# Definimos la lista de páginas HTML que recorreremos para obtener los datos
urlosorno = "http://www.fegosa.cl/preciosciudades/Tosorno.html"
urlpaillaco = "http://www.fegosa.cl/preciosciudades/Tpaillaco.html"
urlptomontt = "http://www.fegosa.cl/preciosciudades/Tptomontt.html"
urlptovaras = "http://www.fegosa.cl/preciosciudades/Tptovaras.html"

newurl = [urlosorno, urlpaillaco, urlptomontt, urlptovaras]

for i in range(4):
   x = newurl[i]
   # Análisis the HTML
   html = urlopen(x)
   soup = BeautifulSoup(html.read(),"html5lib")
   table = soup.find('table')
   rows = table.findAll('tr')
   divs = soup.findAll("table")

   for div in divs:
      rows = div.findAll('tr')
      for row in rows[0:1]:
         column = str(row.findAll('td')[0].text)
         column = column.replace("PRECIOS FEGOSA ", "")
         osorno = column.find("OSORNO")
         paillaco = column.find("PAILLACO")
         montt = column.find("NTT")
         varas = column.find("VARAS")
         if osorno == 4:
            fecha = column.replace("OSORNO", "")
         if paillaco == 4:
            fecha = column.replace("PAILLACO", "")
         if montt == 13:
            fecha = column.replace("PUERTO MONTT", "")
         if varas == 11:
            fecha = column.replace("PUERTO VARAS", "")
      for row in rows[2:]:
         column = str(row.findAll('font')[0].contents)
         column = column.replace("[' ", "")
         column = column.replace("['", "")
         column = column.replace(" ']", "")
         column = column.replace("']", "")
         especie.append(column)
```

```
        date.append(fecha.strip())
        if osorno == 4:
            ciudad.append('Osorno')
        if paillaco == 4:
            ciudad.append('Paillaco')
        if montt == 13:
            ciudad.append('Puerto Montt')
        if varas == 11:
            ciudad.append('Puerto Varas')
    for row in rows[2:]:
        column = str(row.findAll('font')[1].contents)
        column = column.replace("[' ", "")
        column = column.replace("['", "")
        column = column.replace(" ", "")
        column = column.replace(" ']", "")
        column = column.replace("']", "")
        cabezas.append(column)
    for row in rows[2:]:
        column = str(row.findAll('font')[2].contents)
        column = column.replace("[' ", "")
        column = column.replace("['", "")
        column = column.replace(" ", "")
        column = column.replace(" ']", "")
        column = column.replace("']", "")
        pesoProm.append(column)
    for row in rows[2:]:
        column = str(row.findAll('font')[3].contents)
        column = column.replace("[' ", "")
        column = column.replace("['", "")
        column = column.replace(" ", "")
        column = column.replace(" ']", "")
        column = column.replace("']", "")
        primer.append(column)
    for row in rows[2:]:
        column = str(row.findAll('font')[4].contents)
        column = column.replace("[' ", "")
        column = column.replace("['", "")
        column = column.replace(" ", "")
        column = column.replace(" ']", "")
        column = column.replace("']", "")
        segundo.append(column)
    for row in rows[2:]:
        column = str(row.findAll('font')[5].contents)
        column = column.replace("[' ", "")
        column = column.replace("['", "")
        column = column.replace(" ", "")
        column = column.replace(" ']", "")
        column = column.replace("']", "")
        tercero.append(column)
    for row in rows[2:]:
        column = str(row.findAll('font')[6].contents)
        column = column.replace("[' ", "")
        column = column.replace("['", "")
        column = column.replace(" ", "")
        column = column.replace(" ']", "")
        column = column.replace("']", "")
        cuarto.append(column)
    for row in rows[2:]:
        column = str(row.findAll('font')[7].contents)
        column = column.replace("[' ", "")
        column = column.replace("['", "")
        column = column.replace(" ", "")
        column = column.replace(" ']", "")
        column = column.replace("']", "")
        quinto.append(column)
    for row in rows[2:]:
```

```
        column = str(row.findAll('font')[8].contents)
        column = column.replace("[' ", "")
        column = column.replace("['", "")
        column = column.replace(" ", "")
        column = column.replace(" ']", "")
        column = column.replace("']", "")
        promedio.append(column)
    for row in rows[2:]:
        column = str(row.findAll('font')[9].contents)
        column = column.replace("[' ", "")
        column = column.replace("['", "")
        column = column.replace(" ", "")
        column = column.replace(" ']", "")
        column = column.replace("']", "")
        general.append(column)

import pandas as pd

# Definición archivo:
archivo = {'Feria': ciudad, 'Fecha': date, 'Especie': especie, 'Cabezas': cabezas, 'Peso Promedio': pesoProm, '1 er': primer,
'2 do': segundo, '3 er': tercero, '4 to': cuarto, '5 to': quinto, 'Promedio 5PP': promedio, 'Promedio General': general}

# Creación DataFrame:
df_numeros = pd.DataFrame(archivo)

# Guarda datos en CSV:
df_numeros.to_csv('fegosa.csv', header=True, index=False)
```

10. **Dataset. Publicación del dataset en formato CSV en Zenodo con una pequeña descripción.**

    https://zenodo.org/record/3749432

11. **Entrega. Presentar el trabajo con el DOI del dataset en Github.**

    https://github.com/cheukepan/proyectoScraping

12. **Contribuciones**

Contribuciones	Firma
Investigación previa	Diego Cheuquepán Maldonado
Redacción de las respuestas	Diego Cheuquepán Maldonado
Desarrollo código	Diego Cheuquepán Maldonado