Universidad Autónoma de Baja California
Facultad de Ciencias químicas e Ingeniería
Ingeniero en computación

Lenguaje de programación Python
Grupo: 531

TITULO: Proyecto final

NOMBRE: Zavala Roman Irvin Eduardo

PROFESORA: Dra. Karina Raya Díaz

FECHA DE ENTREGA: 05/01/2021

Índice

1. Introducción

El objetivo de este proyecto es usar el lenguaje de programación Python para manipular y mostrar datos de una base de datos, se van a crear: grafos, histogramas, dendrogramas y localización de puntos con el fin de mostrar el comportamiento de estos datos y darles una estética más visual automatizando el proceso.

2. Desarrollo

La base de datos que se utilizará sera la de "Winne Reviews", obtenido de la pagina https://www.kaggle.com/zynicide/wine-reviews ,esta base contiene reseñas de 130,000 vinos alrededor del mundo con calificaciones, comentarios, precios, origen entre otros.

Las columnas de esta base de datos son las siguientes:

Country, description, designation, points, price, province, region_1, region_2, taster_name, taster_twitter_handle, variety, winery.

Las columnas son muy descriptivas con su contenido, los únicos que hacen falta describir son region_1 y region_2; region_1 es la provincia o estado de donde es el viñedo y region_2 no lo tienen todos los vinos, es una región diferente de region_1 pero con viñedos con los que está hecho el vino.

2.1. Grafos e histogramas

En esta sección se va a relacionar países con su promedio de la calificación de vinos, al final se va a juntar por continentes. No se pondrán todos los países del mundo, solo los que se encuentran en la base de datos.

La matriz de adyacencia de países – promedio queda de la siguiente manera:

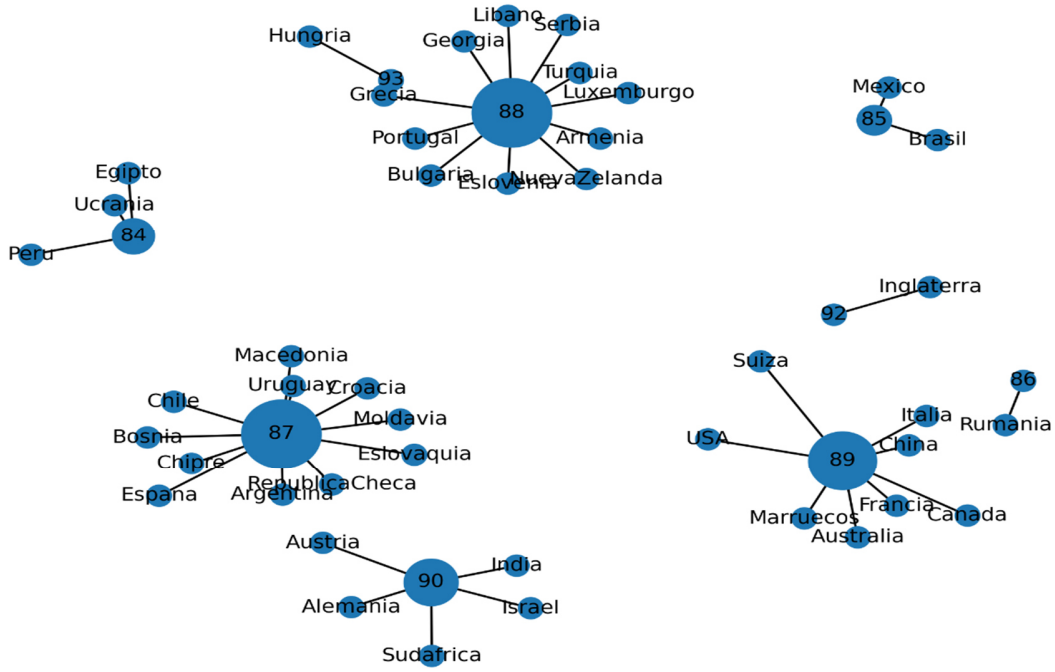
Argentina	87
Armenia	88
Australia	89
Austria	90
Bosnia	87
Brasil	85
Bulgaria	88
Canada	89
Chile	87
China	89
Croacia	87
Chipre	87
Egipto	84
Inglaterra	92
Francia	89
Georgia	88
Alemania	90
Grecia	88
Hungria	93
India	90
Israel	90
Italia	89
Libano	88
Luxemburgo	88
Macedonia	87
Mexico	85
Moldavia	87
Marruecos	89
NuevaZelanda	88
Peru	84
Portugal	88
RepublicaCheca	87
Rumania	86
Serbia	88
Eslovaquia	87

Eslovenia	88
Sudafrica	90
Espana	87
Suiza	89
Turquia	88
Ucrania	84
Uruguay	87
USA	
89	

Con el código del anexo 1 se imprime el siguiente grafo:

Figura 1.

Grafo países – calificaciones.

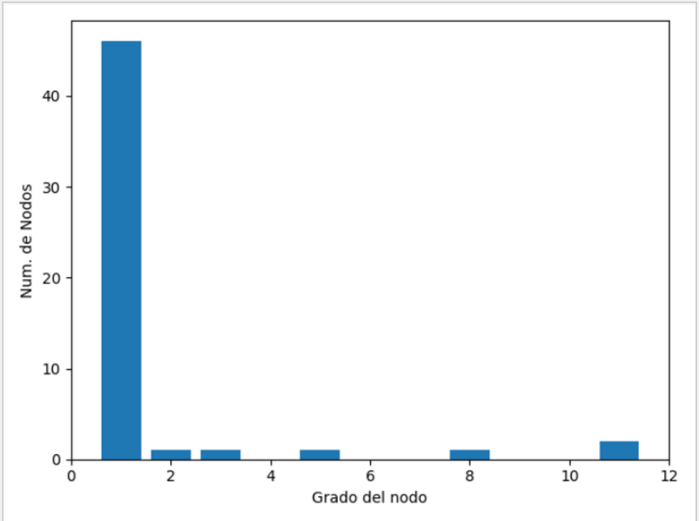


Con este grafo se observa que las calificaciones 87 y 88 se llevan una gran parte de países, el promedio más alto es 93 y solo lo tiene Hungría.

El histograma de grados de nodos impreso por el código es el siguiente:

Figura 2.

Histograma de grados de la figura 1.



Se puede observar que hay más de 40 nodos de grado 1, esto es debido a que todos los países (43) tienen una calificación asociada, por lo que mínimo 43 nodos tendrán grado 1 más los promedios que solo están conectados a un país. Los nodos de grados superiores son promedios: de grado 2 es 85, de grado 3 es 84, de grado 5 es 90, de grado 8 es 89 y por ultimo, 87 y 88 tienen grado 11.

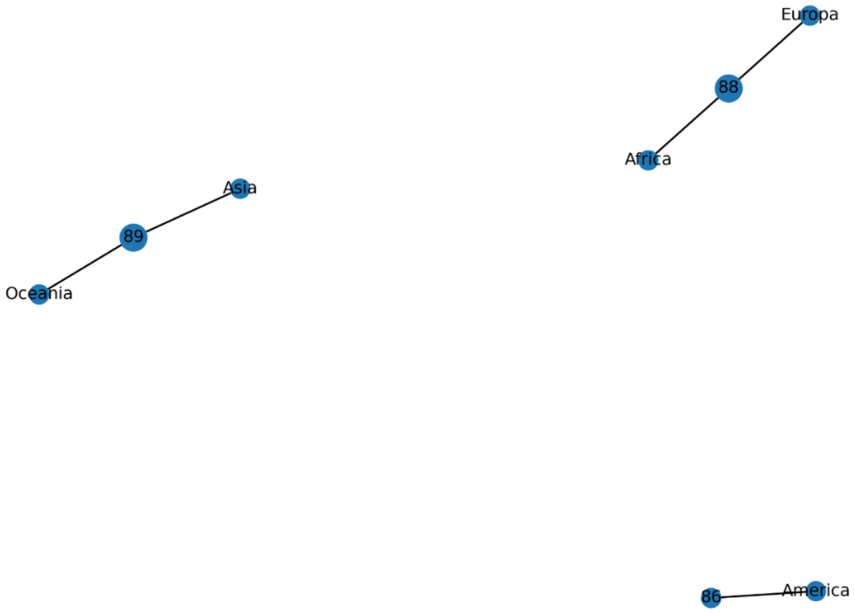
La matriz de adyaciencia continentes – promedio queda así:

America	86
Asia	89
Oceania	89
Africa	88
Europa	88

Con el código del anexo 1 se imprime el siguiente grafo:

Figura 3.

Grafo continentes – calificaciones.

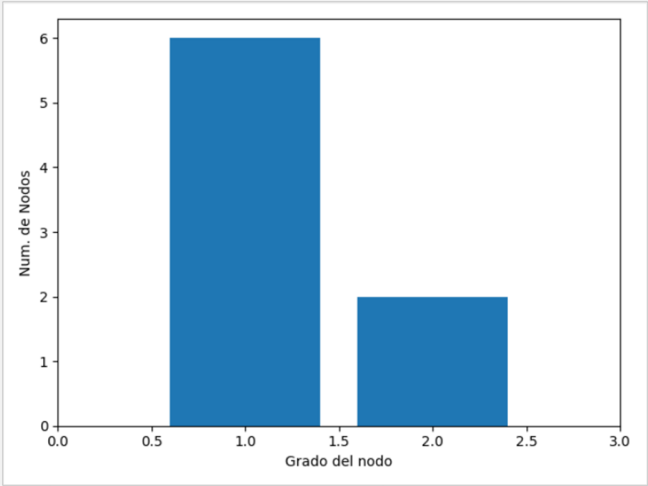


En este grafo se observa que Oceanía y Asia tienen el promedio de vinos más alto en el mundo con 89, el continente Americano se podría decir que tiene el "peor" vino con promedio de 86.

El histograma de grados de nodos impreso por el código es el siguiente:

Figura 4.

Histograma de grados de la figura 3.



Al igual que el grafo anterior, todos los continentes tienen grado 1 y se agregan las calificaciones que solo tienen un país conectado.

Se observa que hay 2 nodos de grado 2, estos son 89 y 88.

2.2. Dendograma

Aunque ya se relaciona países con el promedio de sus vinos, son pocas las variables para saber que países producen los mejores vinos. Para agregar más variables a observar, se va a crear un dendograma de todos los países con el puntaje promedio de los vinos, precio promedio y cantidad de viñedos en el país.

Para saber los promedios basta con la función =PROMEDIO(rango), para la cantidad de viñedos se usó la función =SUMAPRODUCTO(1/CONTAR.SI(rango,rango)). Esta función cuenta las palabras repetidas una por una y da como resultado la cantidad de viñedos en la columna correspondiente.

La matriz de adyaciencia queda de la siguiente manera:

Figura 5.

Matriz de adyaciencia para el dendograma.

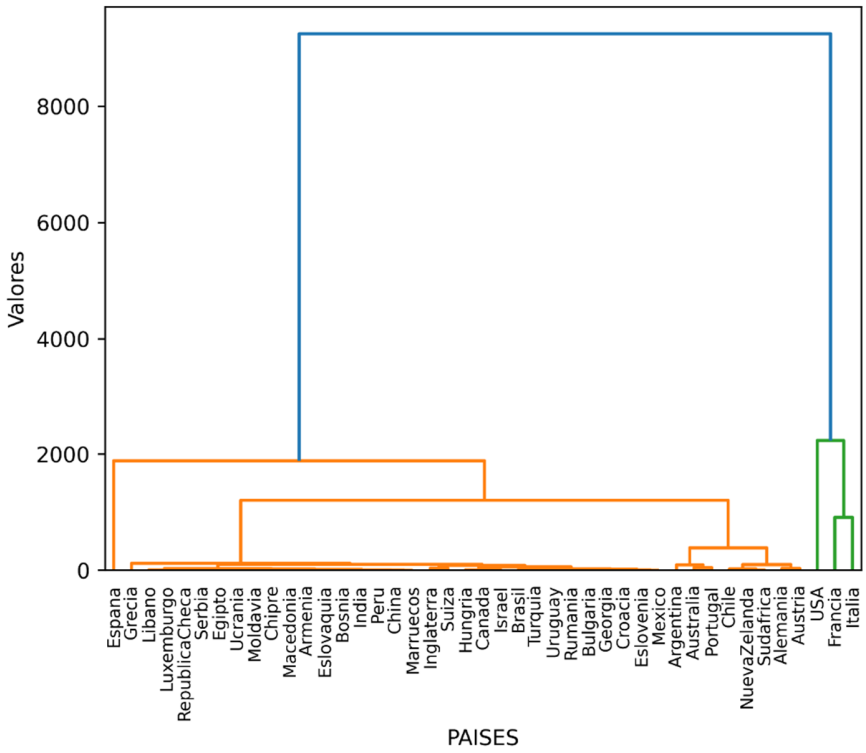
PAIS	CAL PROMEDIO	PRECIO PROMEDIO	CANTIDAD DE VINEDOS
Alemania	90	42.3	256
Argentina	87	36.4	531
Armenia	88	14.5	1
Australia	89	35.3	474
Austria	90	30.8	228
Bosnia	87	12.5	1
Brasil	85	23.8	11
Bulgaria	88	14.6	24
Canada	89	35.7	45
Chile	87	20.8	317
China	89	18	1
Chipre	87	16.3	7
Croacia	87	25.5	32
Egipto	84	0	1
Eslovaquia	87	16	1
Eslovenia	88	24.8	25
Espana	87	28.2	1435

Francia	89	41.1	3844
Georgia	88	19.3	24
Grecia	88	22.3	99
Hungria	93	40.6	41
India	90	13.3	1
Inglaterra	92	51.7	17
Israel	90	31.8	47
Italia	89	39.7	2932
Libano	88	30.7	6
Luxemburgo	88	23.3	2
Macedonia	87	15.6	8
Marruecos	89	19.5	2
Mexico	85	26.8	25
Moldavia	87	16.7	11
NuevaZelanda	88	26.9	300
Peru	84	18.1	2
Portugal	88	26.2	430
RepublicaCheca	87	24.3	3
Rumania	86	15.2	19
Serbia	88	24.5	3
Sudafrica	90	24.7	294
Suiza	89	85.3	4
Turquia	88	24.6	15
Ucrania	84	9.2	4
Uruguay	87	26.4	19
USA	89	36.6	5321

Con el código dado por la maestra el dendograma que se imprime es el siguiente:

Figura 6.

Dendograma de la matriz de adyaciencia de la figura 5.



En este gráfico se puede observar que la isla de países conformado por USA, Francia e Italia en color verde sobresalen comparados con los demás países, si observamos la matriz de adyacencia podemos ver que tienen una cantidad de viñedos mucho mayor a los demás países, con un mínimo de 2932 en Italia. En la otra isla de países en anaranjado España esta sobresale de todos los demás países anaranjados unidos, esto se debe a su cantidad de viñedos que sobresale por mucho de los demás pero sin llegar a los países en verde.

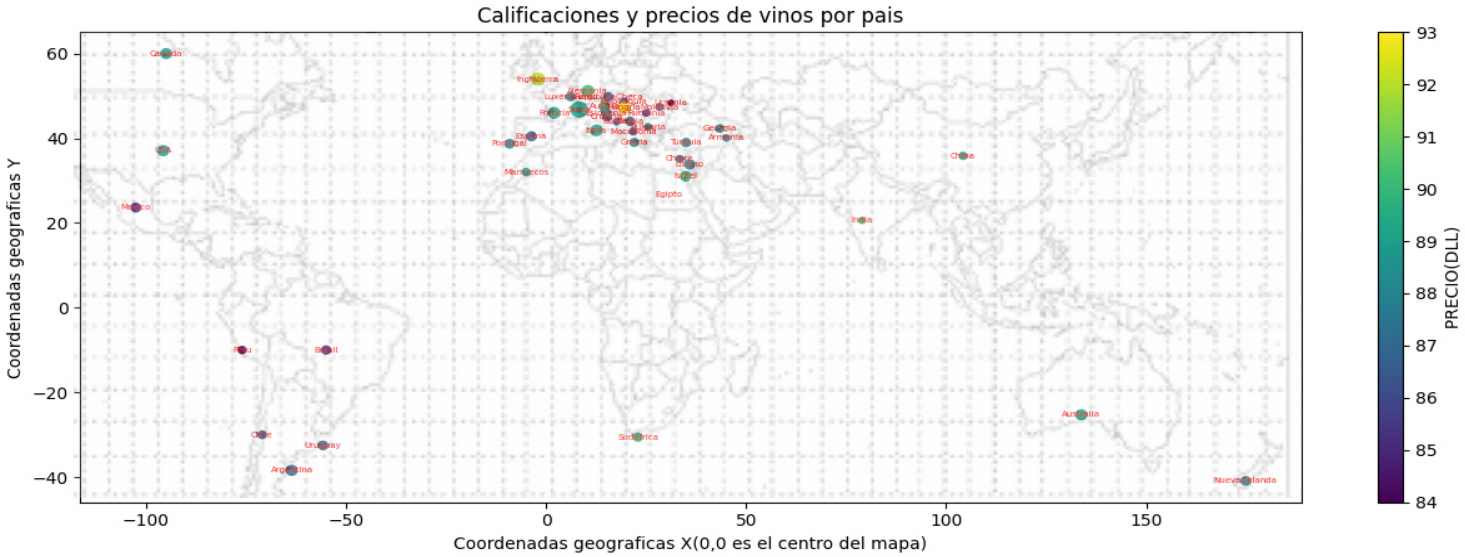
2.3. Geolocalizacion de vinos en el mapa

Para mostrar el precio y promedio de una manera más intuitiva se mostrará un mapa con nodos, estos nodos representan los países en su posición en el mapa con un tamaño y color. El tamaño significa el promedio del vino en el país y su color el precio. Para poder dar una posición en un mapa hay que poner coordenadas en cada nodo, lo más sencillo son las coordenadas decimales y se pondrá en un mapa de -180 a 180.

El gráfico impreso es el siguiente:

Figura 5.

Geolocalizacion de los vinos con precio y calificacion.



El mapa de fondo fue un poco difícil de encajar con los datos, en sí el poner la imagen de fondo no fue complicado, pero hacer que coincidiera con los ejes fue más de prueba y error con el código añadido.

Los cambios al código con el dado por la maestra fue lo siguiente:

-Cambio de contenido en los G.add_node
-Uso de axes_coord, ax_image y otras herramientas para insertar el mapa de fondo

3. Anexos

Anexo 1.

Código para los grafos e histogramas.

```
#!/usr/bin/python
# -*- coding: utf-8 -*-
#from pylab import figure
import pylab
import networkx as nx
import matplotlib.pyplot as plt


G=nx.read_weighted_edgelist('Continentes promedio.txt', create_using=nx.MultiGraph())#Este es el grafo de continentes-promedios
F=nx.read_weighted_edgelist('Valores redondeados vinitos.txt', create_using=nx. MultiGraph ())#Este es el grafo de paises-promedios

custom_node_color={}
custom_node_color['0'] = 'black'

#El codigo se ejecuta una vez por grafo, sino se sobreponen. No pude automatizarlo :(

#nx.draw(G,pos = nx.spring_layout(G,k=0.17,iterations=20), node_size = [G.degree(i) * 100 for i in G.nodes()], with_labels = True, font_size = 9)
#plt.savefig("G.png", dpi=300)
#nx.draw(F,pos = nx.spring_layout(F,k=0.17,iterations=20), node_size = [F.degree(i) * 100 for i in F.nodes()], with_labels = True, font_size = 9)
#plt.savefig("F.png", dpi=300)


#Archivo con el dato del promedio del grado de clusterizacion por nodo
avg_node_degree = nx.average_neighbor_degree(G)
vString = str(avg_node_degree)

def make_histogram(aGraph):
    fig = pylab.figure()
    hist = nx.degree_histogram(aGraph)
    pylab.bar(range(len(hist)), hist, align = 'center')
    pylab.xlim((0, len(hist)))
    pylab.xlabel("Grado del nodo")
    pylab.ylabel("Num. de Nodos")
    return fig
def save_histogram(aGraph, filename):
    fig = make_histogram(aGraph)
    fig.savefig(filename)
```

```
save_histogram(G,"CONTINENTES promedio.png")
pylab.show()
save_histogram(F,"CALIFICACIONES.png")
pylab.show()
```

Anexo 2.

Código para el dendograma.

```
# -*- coding: utf-8 -*-
import pandas as pd
from matplotlib import pyplot as plt
from scipy.cluster import hierarchy
import numpy as np

df = pd.read_csv("DENDOGRAMA VINITOS.csv")
df = df.set_index('PAIS')
df = df.rename_axis(None, axis = 1)#Tuve que cambiar el del porque me daba error


# Calculate the distance between each sample
Z = hierarchy.linkage(df, 'ward')
print(Z)

# Plot with Custom leaves
hierarchy.dendrogram(Z, leaf_rotation=90, leaf_font_size=8, labels=df.index)
plt.axis('on')
# plt.title('Clasificacion de la redes respecto a su uso')
plt.xlabel('PAISES')
plt.ylabel('Valores')
plt.savefig("DENDOGRAMA PAISES.png", dpi=600, bbox_inches='tight')


plt.show()
```

Anexo 3.

Código para geolocalización de puntos.

```
import networkx as nx
import matplotlib.pyplot as plt

G = nx.Graph()

#País -> Coordenadas decimales del pais -> Promedio -> Precio(dll)
G.add_node('Alemania', pos = ( 10.451526,51.165691    ), promedio =  90 , precio =    42.3   )
G.add_node('Argentina', pos = (      -63.616672, -38.416097     ), promedio =  87 , precio =    36.4   )
G.add_node('Armenia', pos = (   45.038189, 40.069099  ), promedio =  88 , precio =    14.5   )
G.add_node('Australia', pos = (    133.775136, -25.274398    ), promedio =  89 , precio =    35.3   )
G.add_node('Austria', pos = (   14.550072, 47.516231  ), promedio =  90 , precio =    30.8   )
G.add_node('Bosnia', pos = (    17.679076, 43.915886  ), promedio =  87 , precio =    12.5   )
G.add_node('Brasil', pos = (   -55,-10    ), promedio =  85 , precio =    23.8   )
G.add_node('Bulgaria', pos = ( 25.48583,42.733883 ), promedio =  88 , precio =    14.6   )
```

```
G.add_node('Canada', pos = (  -95.0000000, 60.0000000   ), promedio =  89 , precio =    35.7   )
G.add_node('Chile', pos = (   -71.0000000, -30   ), promedio =  87 , precio =    20.8   )
G.add_node('China', pos = (   104.195397, 35.86166   ), promedio =  89 , precio =    18 )
G.add_node('Chipre', pos = (  33.429859, 35.126413   ), promedio =  87 , precio =    16.3   )
G.add_node('Croacia', pos = ( 15.2, 45.1 ), promedio =  87 , precio =    25.5   )
G.add_node('Egipto', pos = (  30.802498, 26.820553   ), promedio =  84 , precio =    0   )
G.add_node('Eslovaquia', pos = (  19.5000000, 48.6667000 ), promedio =  87 , precio =    16 )
G.add_node('Eslovenia', pos = (   15.0000000, 46 ), promedio =  88 , precio =    24.8   )
G.add_node('Espana', pos = (  -3.7, 40.46    ), promedio =  87 , precio =    28.2   )
G.add_node('Francia', pos = (  2.0000000, 46   ), promedio =  89 , precio =    41.1   )
G.add_node('Georgia', pos = (  43.356892, 42.315407   ), promedio =  88 , precio =    19.3   )
G.add_node('Grecia', pos = (   22.0000000, 39 ), promedio =  88 , precio =    22.3   )
G.add_node('Hungria', pos = (  19.503304,47.162494    ), promedio =  93 , precio =    40.6   )
G.add_node('India', pos = (   78.96288, 20.593684    ), promedio =  90 , precio =    13.3   )
G.add_node('Inglaterra', pos = (  -2, 54 ), promedio =  92 , precio =    51.7   )
G.add_node('Israel', pos = (  34.851612, 31.046051   ), promedio =  90 , precio =    31.8   )
G.add_node('Italia', pos = (  12.56738, 41.87194 ), promedio =  89 , precio =    39.7   )
G.add_node('Libano', pos = (  35.862285, 33.854721   ), promedio =  88 , precio =    30.7   )
G.add_node('Luxemburgo', pos = (  6.129583, 49.815273    ), promedio =  88 , precio =    23.3   )
G.add_node('Macedonia', pos = (   21.745275, 41.608635   ), promedio =  87 , precio =    15.6   )
G.add_node('Marruecos', pos = (   -5.0000000, 32 ), promedio =  89 , precio =    19.5   )
G.add_node('Mexico', pos = (  -102.552784,23.634501   ), promedio =  85 , precio =    26.8   )
G.add_node('Moldavia', pos = ( 28.369885, 47.411631   ), promedio =  87 , precio =    16.7   )
G.add_node('NuevaZelanda', pos = ( 174.885971, -40.900557 ), promedio =  88 , precio =    26.9   )
G.add_node('Peru', pos = ( -76, -10   ), promedio =  84 , precio =    18.1   )
G.add_node('Portugal', pos = ( -9.13333, 38.71667 ), promedio =  88 , precio =    26.2   )
G.add_node('RepublicaCheca', pos = (  15.5000000, 49.7500000 ), promedio =  87 , precio =    24.3   )
G.add_node('Rumania', pos = (  25, 46 ), promedio =  86 , precio =    15.2   )
G.add_node('Serbia', pos = (  21, 44 ), promedio =  88 , precio =    24.5   )
G.add_node('Sudafrica', pos = (   22.937506, -30.559482   ), promedio =  90 , precio =    24.7   )
G.add_node('Suiza', pos = (   8.227512, 46.818188    ), promedio =  89 , precio =    85.3   )
G.add_node('Turquia', pos = (  35, 39 ), promedio =  88 , precio =    24.6   )
G.add_node('Ucrania', pos = (  31.16558, 48.379433   ), promedio =  84 , precio =    9.2   )
G.add_node('Uruguay', pos = (  -55.765835, -32.522779 ), promedio =  87 , precio =    26.4   )
G.add_node('USA', pos = (  -95.712891, 37.09024   ), promedio =  89 , precio =    36.6   )


pos = nx.get_node_attributes(G, 'pos')
promedio = nx.get_node_attributes(G,'promedio')
precio = nx.get_node_attributes(G,'precio')


G.nodes(data=True)

#Elegi que el tamano de las bolitas sea el precio del vino por país y su color el promedio
node_color=[float(promedio[v]) for v in G]#El color es la calificacion
node_size=[float(precio[v]) for v in G]#El tamano es el precio


cmap = plt.cm.viridis#Cambie de rango de color para adaptar al que tomaba por defecto en titulos

fig, ax = plt.subplots(figsize=(13.1 ,5))#Para los ejes
#Titulo y eje x/y
plt.title('Calificaciones y precios de vinos por pais')
plt.xlabel('Coordenadas geograficas X(0,0 es el centro del mapa)')
plt.ylabel('Coordenadas geograficas Y')
```

```
nx.draw(G, pos, font_size=5, font_color = "red", with_labels=True, node_size=node_size,
node_color=node_color, edge_cmap=cmap, ax=ax)

sm = plt.cm.ScalarMappable(cmap=cmap, norm=plt.Normalize(84,93))#Cambie el parametro de normalize
porque daba error con node_color
ax.set_axis_on()
ax.tick_params(left=True, bottom=True, labelleft=True, labelbottom=True)
sm._A = []
a = plt.colorbar(sm)
a.set_label('                              PROMEDIO                              ')

#Para insertar el mapa, axes coord. sirve para mover el mapa de fondo por el plot
axes_coords = [0.052, 0, 0.75, 1.05]
ax_image = plt.gcf().add_axes(axes_coords)
img = plt.imread('unnamed_3.jpg')
ax_image.imshow(img, zorder = 0, alpha = 0.2) #Alpha para dar transparencia al mapa
ax_image.axis('off')

plt.show()
```

4. Conclusiones

Los 4 maneras de mostrar la base de datos son útiles dependiendo de nuestro propósito, los grafos y nodos muestran las uniones entre los datos, dándonos información relevante como que las calificaciones 88 y 89 son las más comunes entre todos los países. El dendograma al usar más variables sirve para mostrar los países que tienen los números más altos y separarlos por grupos. Por último, la localización de puntos en el mapa en mi opinión es la más útil y atractiva visualmente, muestra los vinos en su país diciendo el precio y calificación del vino con colores y tamaños, siendo más accesible a gente que no tiene ni idea de mi base de datos.