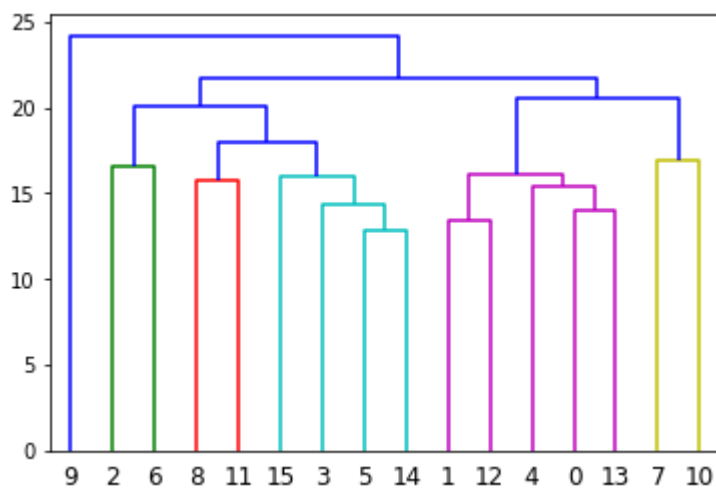
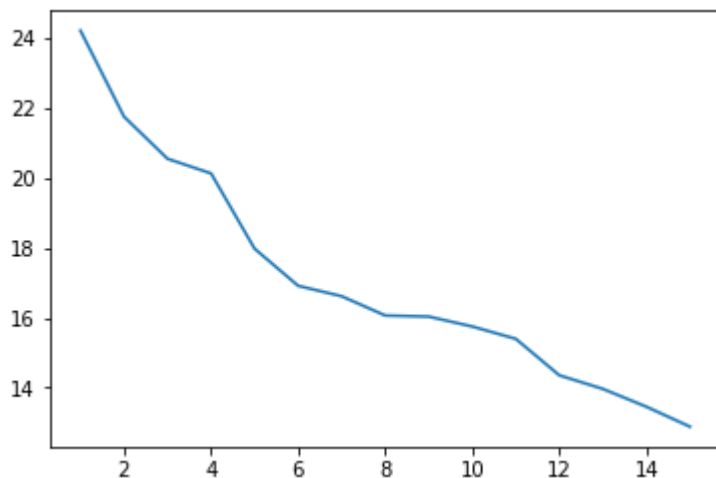


## Hierarchical clustering.

Se analiza la base de datos no estandarizada (debido a que el rango en el cuál se obtienen los datos va de 0 a 5 en todos los casos) con hierarchical clustering en método completo y utilizando la métrica de distancia euclidiana (debido a que son datos ordinales, dónde una calificación de 1 es menor a 5). Los resultados obtenidos se pueden representar en el diagrama siguiente:



Cada uno de los números de abajo es el 'ID' de los usuarios clasificados. Se puede apreciar que existen dos grandes grupos y un compañero 'radical'; el 9 se encuentra alejado de todos, del 2 al 14 (en orden del dendrograma) representarían el primer grupo y del 1 al 10 el segundo grupo. Si se desearan más particiones podríamos llegar a aquellas que se encuentran de colores, en ese punto tendríamos 6 grupos similares, dónde el más grande es el morado.



Se busca confirmar que los grupos están correctamente formados, para ello hacemos una gráfica de codos y encontramos que existe un posible codo en 3 y 6 (como lo habíamos notado visualmente antes de hacer algún calculo). A pesar de la existencia de los codos; la diferencia entre cada uno de los usuarios es bastante grande.

En la imagen a la derecha se analiza la pertenencia del usuario en los grupos. En este caso Oscar Flores (usuario 1) pertenece al grupo 4, por lo tanto, usuarios 0, 4, 12 y 13 deberían tener gustos similares.

Una vez que conocemos los usuarios similares observamos las películas de los 5.

American Pie	Los cuatro los nobes	La máscara	señor de los anillos	Harry Potter	Blanca Nieves	Big fish	Narnia	laberinto del faun	The shape of water	Aquaman	lobo de wall street	de eventos desast
1	5	2	2	5	3	1	5	3	1	1	5	1
4	5	1	4	3	3	1	3	1	4	1	5	1
5	5	4	2	2	3	3	4	1	1	1	5	1
4	3	5	5	4	2	3	4	3	1	1	5	2
4	4	2	4	5	2	1	4	5	4	1	4	1

Your name	Coco	Intensa mente	El viaje de Shiro	ero 6 (Grandes he	Shrek	Toy story	Lilo y Stitch	Spider-Man	Fantastic Mr. Fox	South park	Trolls	si entrenar a tu di
1	2	1	5	1	5	5	5	1	3	1	5	5
1	3	4	1	5	3	3	3	1	1	1	1	4
1	3	1	1	1	5	5	3	5	1	5	1	4
1	1	1	1	1	2	5	1	3	1	5	1	5
1	4	5	1	1	5	4	4	5	1	3	1	3

V de Vendetta	Butter	Contacto	Origen (Inception)	Memento	I Origins	Lucy	La terminal	una pasión (The r	la sombra del an	metros sobre el ci	The Now	Titanic
2	4	1	5	5	1	1	2	5	1	1	1	3
5	5	1	5	4	1	1	3	3	1	1	1	3
1	3	1	1	1	1	1	1	3	1	3	1	4
4	3	1	1	4	1	1	1	1	1	3	1	4
3	5	1	4	3	1	4	1	1	1	4	1	4

	0
0	4
1	4
2	1
3	3
4	4
5	3
6	1
7	5
8	2
9	6
10	5
11	2
12	4
13	4
14	3
15	3

Se puede notar que existen algunas películas en dónde la mayoría coincide. No obstante, hay otras en las que ninguno entra en acuerdo con los demás usuarios, se cree que esta disparidad ocurre debido a que algún usuario no vio alguna película (y la calificó con 1), mientras que los demás le dieron puntuaciones altas o porque estamos forzando al sistema a hacer grupos en dónde no existen usuarios realmente similares (muy pocos datos).

Código:

```
8 import pandas as pd
9 import numpy as np
10 import scipy.cluster.hierarchy as hi
11 import matplotlib.pyplot as plt
12
13 ### Seleccionar datos
14 data = pd.read_excel('../data/Test de películas(1-16).xlsx')
15
16 csel = np.arange(6,243,3)
17 cnames = list(data.columns.values[csel])
18 datan = data[cnames]
19
20 ### Dendrograma
21 Z = hi.linkage(datan,metric='euclidean',method='complete')
22
23 ### Grafica de dendrograma
24 hi.dendrogram(Z)
25
26 ### Grafica de codo
27 last = Z[:,2]
28 last = last[::-1]
29 plt.plot(np.arange(len(last))+1,last)
30
31 ### Usuario 1: Oscar Flores
32 sim = hi.fcluster(Z,6,criterion='maxclust') #pertenezco al grupo 4
33 pel_sim = datan[sim==4]
```