Que harían en los casos que les de valores Nan? Porque les puede dar el valor Nan. Realice un análisis. ¿Cuál es su propuesta de solución?

Analizando la correlación de pearson

$$r = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sqrt{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}} \sqrt{\sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n}}}$$

r =	#¡DIV/0!	0/0

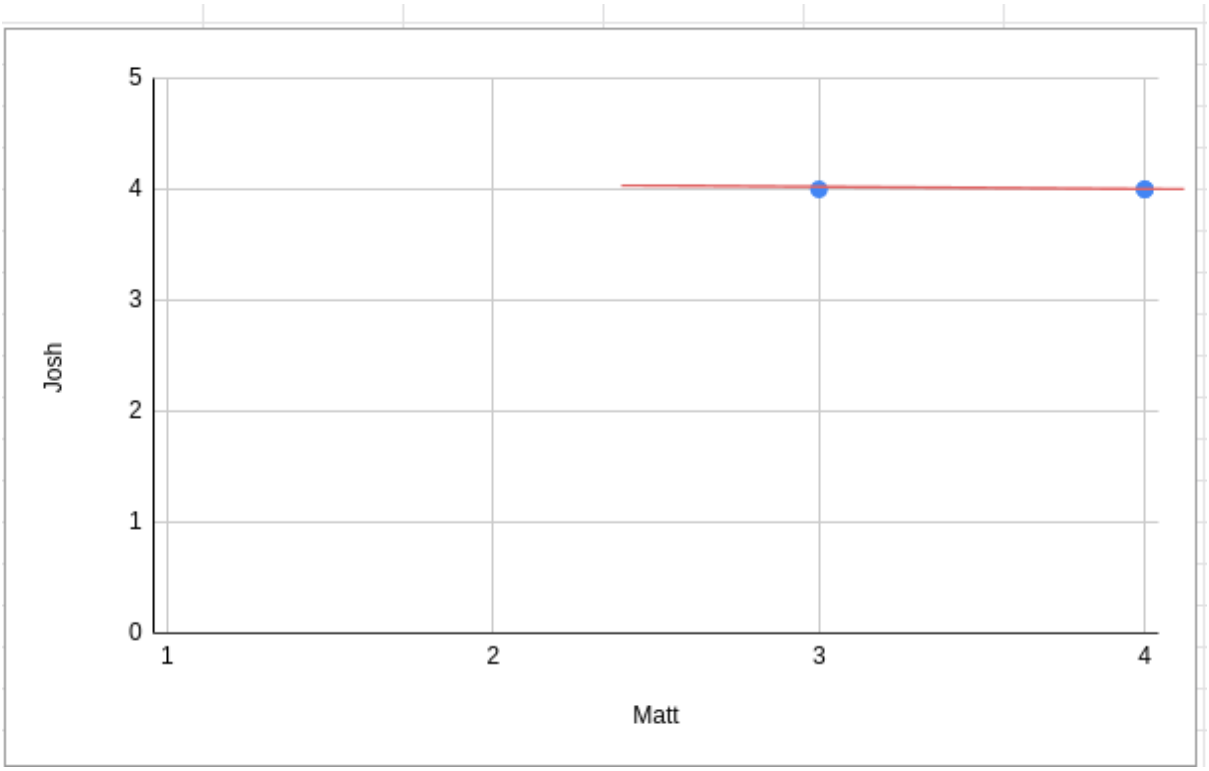
El valor nan devuelta por python es cuando el numerador y el denominador son ceros (0 / 0) pero el resultado teóricamente deberían ser indeterminado por lo tanto es difícil medir la correlación entre las variables y por ello no podemos demostrar si es de acuerdo perfecto o no.

Datos:

Como se muestra en la imagen podemos observar que la cantidad de celdas que no están calificadas mayor que celdas con calificación ya que solo 3 películas están con su calificación de cada usuario.

Según la grafica de los puntos en plano con respecto a la correlación se nota que la varianza de y es nula y por es que la correlación es indeterminada

	Matt (X)	Josh (Y)
Alien		3
Avatar		4
Blade Runner		
Braveheart		
Dodgeball		
Forest Gump	4	4
Gladiator		4
Jaws		5
Kazaam		
Lord of the Rings	1	
Napolean Dynamite	3	
Old School		
Pootie Tang		
Pulp Fiction		
Scarface		
Shawshank Redemption		
Snakes on a Plane	1	
Spiderman	3	4
Star Wars		5
The Dark Knight		4
The Happening		
The Matrix		5
Toy Story	4	4
Village		4
You Got Mail		
n=		3



Los datos son dispersos y no indican si Matt y Josh tienen gustos similares de las 25 películas por ello se ha tratado de comprobar que pasaría si los valores que Matt no califica pero si Josh o que Josh califica pero no Matt se establecería con cero de la siguiente manera

Matt (X)	Josh (Y)
0	3
0	4
4	4
0	4
0	5
1	0
3	0
1	0
3	4
0	5
0	4
0	5
4	4
0	4
	14

$$r = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sqrt{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}} \sqrt{\sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n}}}$$

r =	-0,2204095586

Notablemente la cantidad de datos tomados serían 14 y el resultado de la ecuación de la correlación de Pearson cambiaria con un valor que podemos usar para determinar cuán asociadas son las variables entre sí.

En este caso el resultado es menor a 0 esto quiere decir que las variables se relacionan inversamente. Pero al ser datos de calificaciones que van de 0 a 5, no podemos asegurar que realmente sea así, ya que se alteró los valores de datos con cero por ello es como si el usuario no le agrada dicha película llegando a una conclusión de que los datos alterados son ruidos.

Si en las 3 calificaciones notamos que tienen gustos similares y mediante ello se podría rellenar los datos faltantes con gusto similar el resultado de la correlación sería mayor a 0. Pero también se genera ruido.

Que harían en los casos que les de valores Nan?

Para estos casos es mejor usar otras métricas de distancia que pueden ser euclidean, manhattan o aquella métricas que se comportan mejor con datos esparsos como es la similaridad del coseno.

En este caso si observamos los datos vemos que la calificación de las tres películas son similares y se podría decir que tiene una relación directa.

Como se muestra en la figura siguiente notamos que Matt y Josh tienen gustos similares

$$\cos(x,y) = \frac{x \cdot y}{\|x\| \times \|y\|}$$

cos(X,Y)	0,9918365981

¿Cuál es su propuesta de solución?

Debido a la cantidad de calificaciones no podemos determinar si Matt o Josh gustan o disgustan del resto de las películas para poder alterar los datos

Por lo tanto sería aumentar mayormente la cantidad de datos según los 3 datos similares para así evitar valores indeterminados.