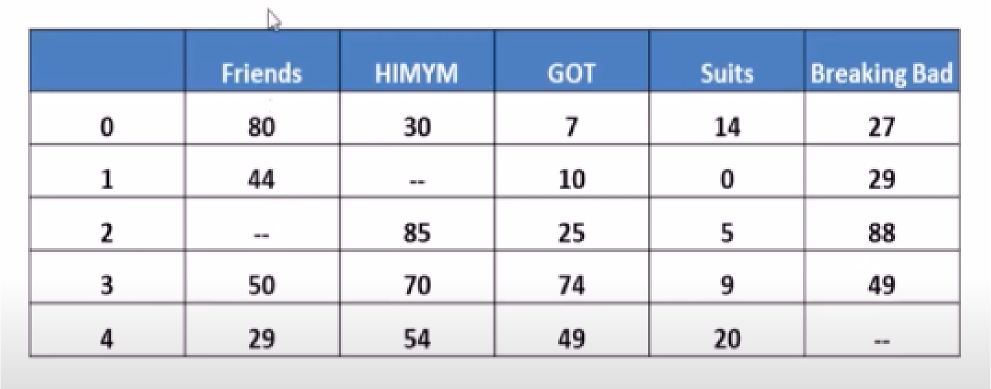
KNN IMPUTATION FOR MISSING VALUES

It is the multivariate imputation

DATASET:



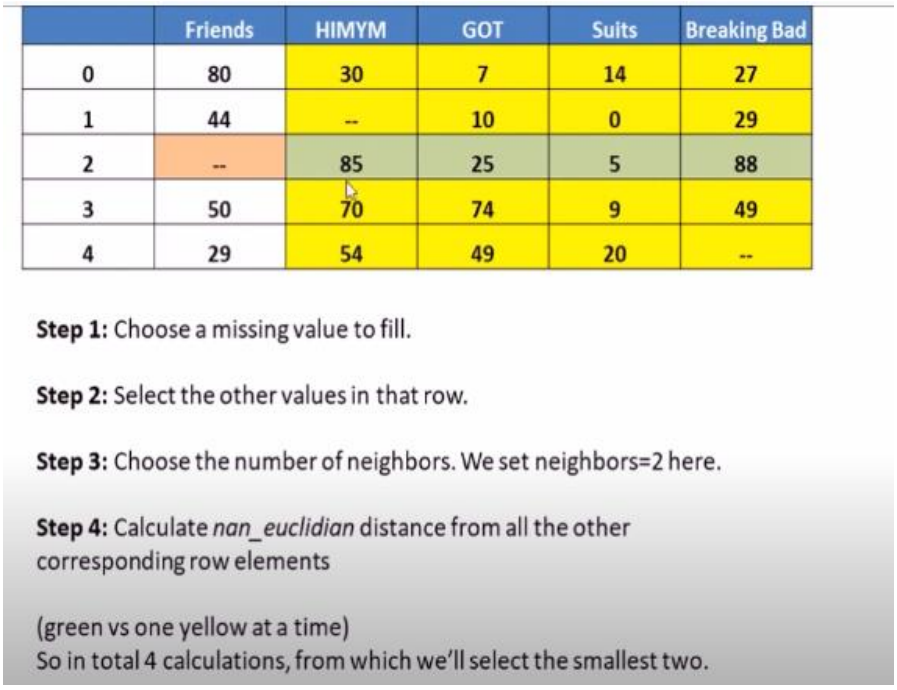
	Friends	HIMYM	GOT	Suits	Breaking Bad
0	80	30	7	14	27
1	44	--	10	0	29
2	--	85	25	5	88
3	50	70	74	9	49
4	29	54	49	20	--

In this dataset we can able to find null values in all the columns. So we want to use KNN imputation

The missing values will be calculated by taking the values of the neighbour of the particular value.

The column with null value will considered as the output feature at that instant.

The null value will be calculated using Euclidian distances



	Friends	HIMYM	GOT	Suits	Breaking Bad
0	80	30	7	14	27
1	44	--	10	0	29
2	--	85	25	5	88
3	50	70	74	9	49
4	29	54	49	20	--

Step 1: Choose a missing value to fill.

Step 2: Select the other values in that row.

Step 3: Choose the number of neighbors. We set neighbors=2 here.

Step 4: Calculate *nan_euclidian* distance from all the other corresponding row elements

(green vs one yellow at a time)
So in total 4 calculations, from which we'll select the smallest two.

The K represents number of neighbours in KNN.

In this calculation 4 neighbours has been took into reference.

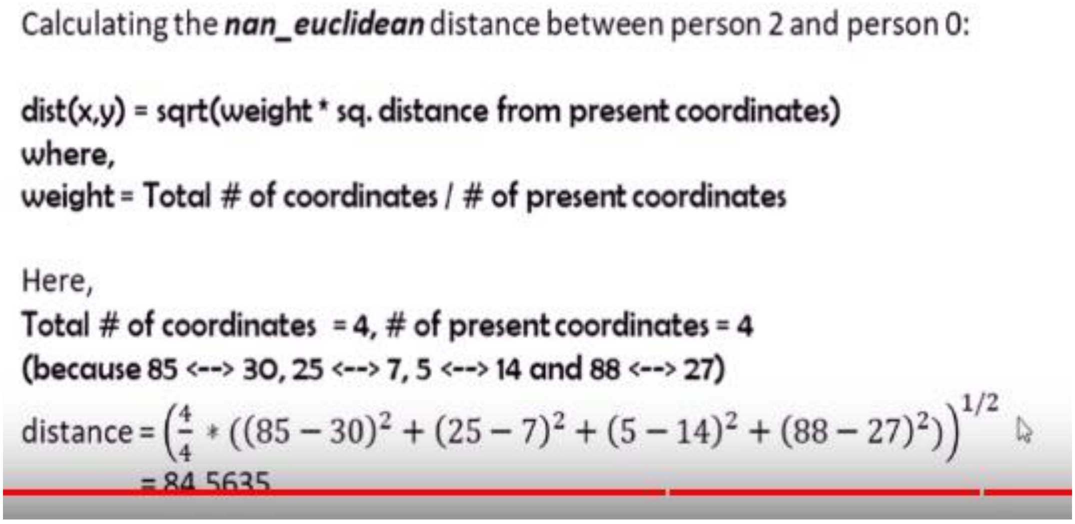
CALCULATION:-

Total # of coordinates means no of columns considered excluding the column of missing values.

Total # present coordinates means during a calculation between a single neighbour row in the instance and row containing missing value.Both the rows should contain value.

If both missing row and neighbour row has a value then it is a present coordinate.

If any one of the value missed in row then it is not present coordinate.



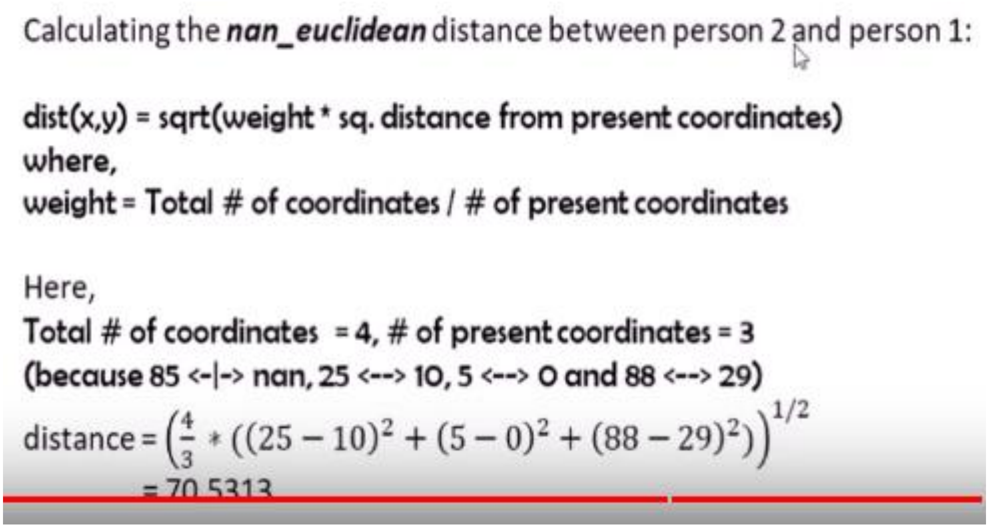
Calculating the ***nan_euclidean*** distance between person 2 and person 0:

dist(x,y) = sqrt(weight * sq. distance from present coordinates)
where,
weight = Total # of coordinates / # of present coordinates

Here,
Total # of coordinates = 4, # of present coordinates = 4
(because 85 <--> 30, 25 <--> 7, 5 <--> 14 and 88 <--> 27)

$$\text{distance} = \left(\frac{4}{4} * \left((85-30)^2 + (25-7)^2 + (5-14)^2 + (88-27)^2\right)\right)^{1/2}$$
$$= 84.5635$$

Here total # of coordinates and present coordinates are 4 because in both the rows that is the row of person 0 and person 2 there is no missing value.(excluding missing feature)



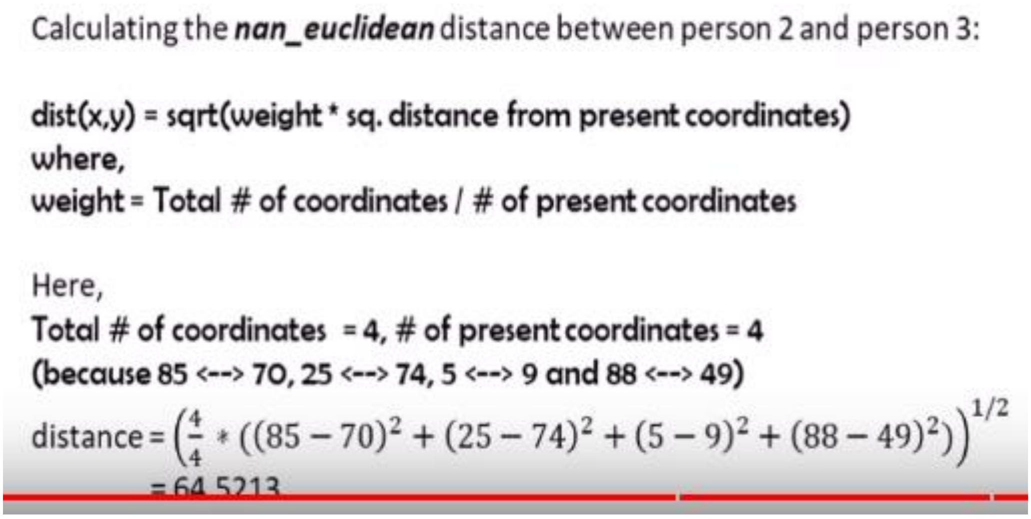
Calculating the ***nan_euclidean*** distance between person 2 and person 1:

dist(x,y) = sqrt(weight * sq. distance from present coordinates)
where,
weight = Total # of coordinates / # of present coordinates

Here,
Total # of coordinates = 4, # of present coordinates = 3
(because 85 <-|-> nan, 25 <--> 10, 5 <--> 0 and 88 <--> 29)

distance = $\left(\frac{4}{3} * \left((25-10)^2 + (5-0)^2 + (88-29)^2\right)\right)^{1/2}$
= 70.5313

Here present coordinates is 3 because there is a missing values in himym column. So we cant take that into reference.

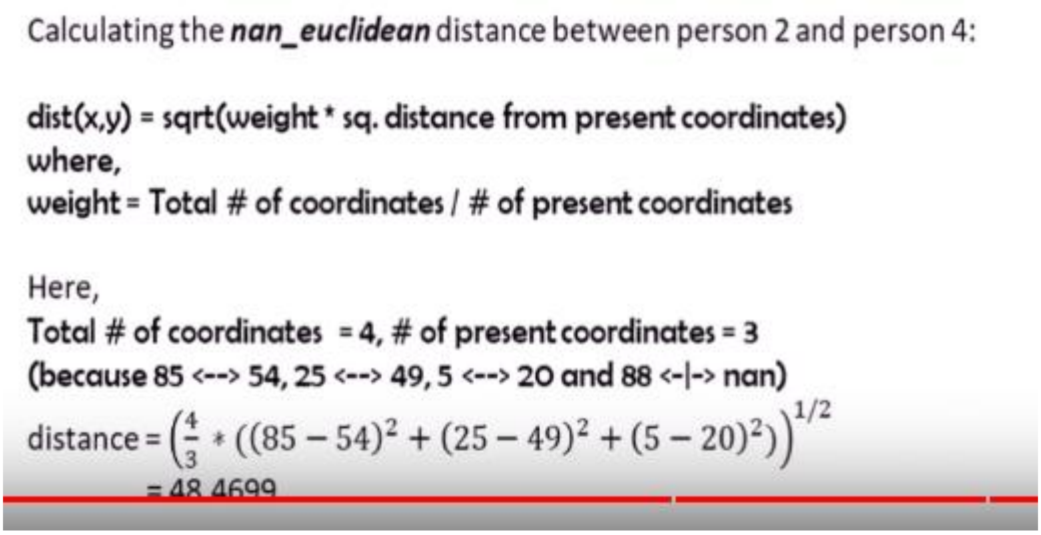


Calculating the ***nan_euclidean*** distance between person 2 and person 3:

dist(x,y) = sqrt(weight * sq. distance from present coordinates)
where,
weight = Total # of coordinates / # of present coordinates

Here,
Total # of coordinates = 4, # of present coordinates = 4
(because 85 <--> 70, 25 <--> 74, 5 <--> 9 and 88 <--> 49)

distance = $\left(\frac{4}{4} * \left((85-70)^2 + (25-74)^2 + (5-9)^2 + (88-49)^2\right)\right)^{1/2}$
= 64.5213

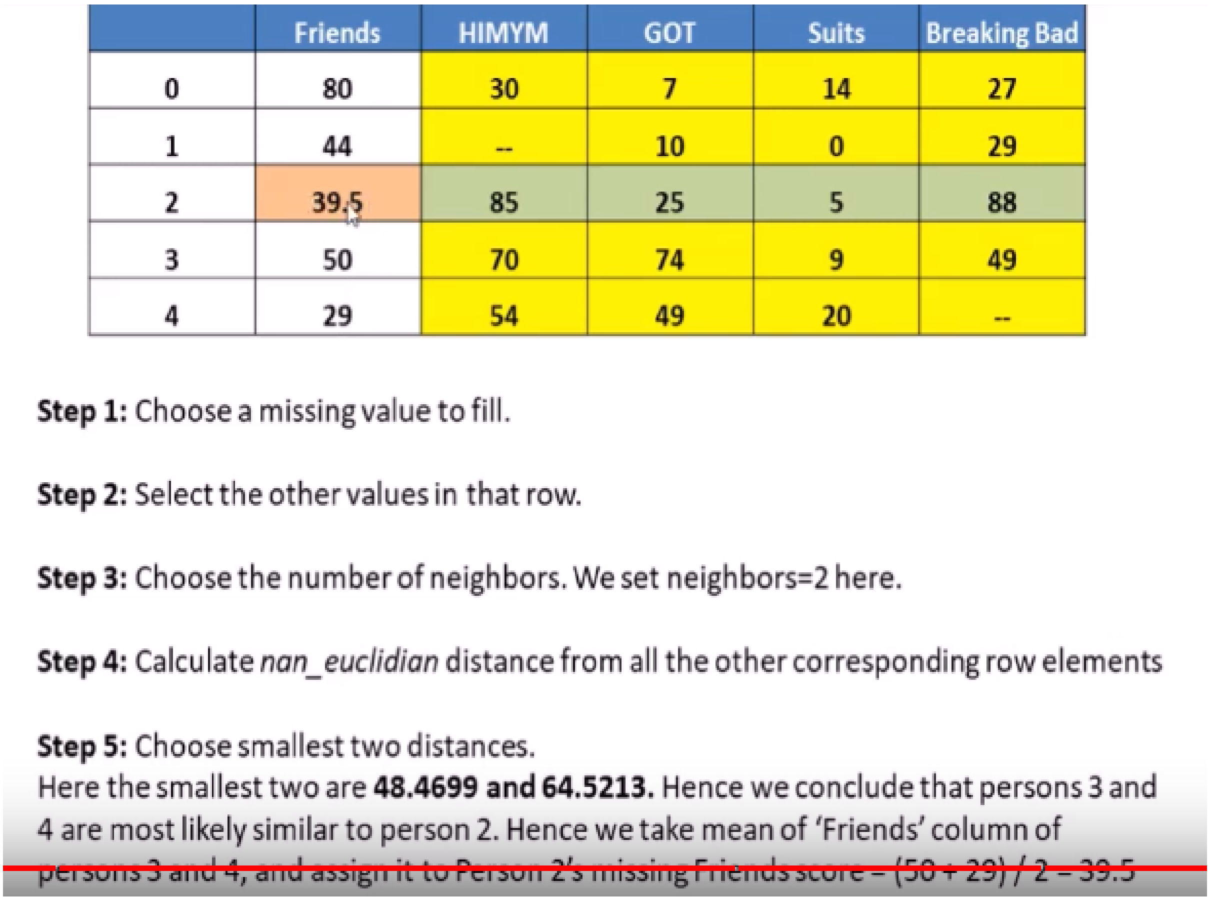


Calculating the ***nan_euclidean*** distance between person 2 and person 4:

dist(x,y) = sqrt(weight * sq. distance from present coordinates)
where,
weight = Total # of coordinates / # of present coordinates

Here,
Total # of coordinates = 4, # of present coordinates = 3
(because 85 <--> 54, 25 <--> 49, 5 <--> 20 and 88 <-|-> nan)

distance = $\left(\frac{4}{3} * \left((85-54)^2 + (25-49)^2 + (5-20)^2\right)\right)^{1/2}$
= 48.4699



	Friends	HIMYM	GOT	Suits	Breaking Bad
0	80	30	7	14	27
1	44	--	10	0	29
2	39.5	85	25	5	88
3	50	70	74	9	49
4	29	54	49	20	--

Step 1: Choose a missing value to fill.

Step 2: Select the other values in that row.

Step 3: Choose the number of neighbors. We set neighbors=2 here.

Step 4: Calculate *nan_euclidian* distance from all the other corresponding row elements

Step 5: Choose smallest two distances.
Here the smallest two are **48.4699 and 64.5213.** Hence we conclude that persons 3 and 4 are most likely similar to person 2. Hence we take mean of 'Friends' column of persons 3 and 4, and assign it to Person 2's missing Friends score = (50 + 29) / 2 = 39.5

Like wise we want to do it for all the values.

Scikit learn has the KNN imputer.