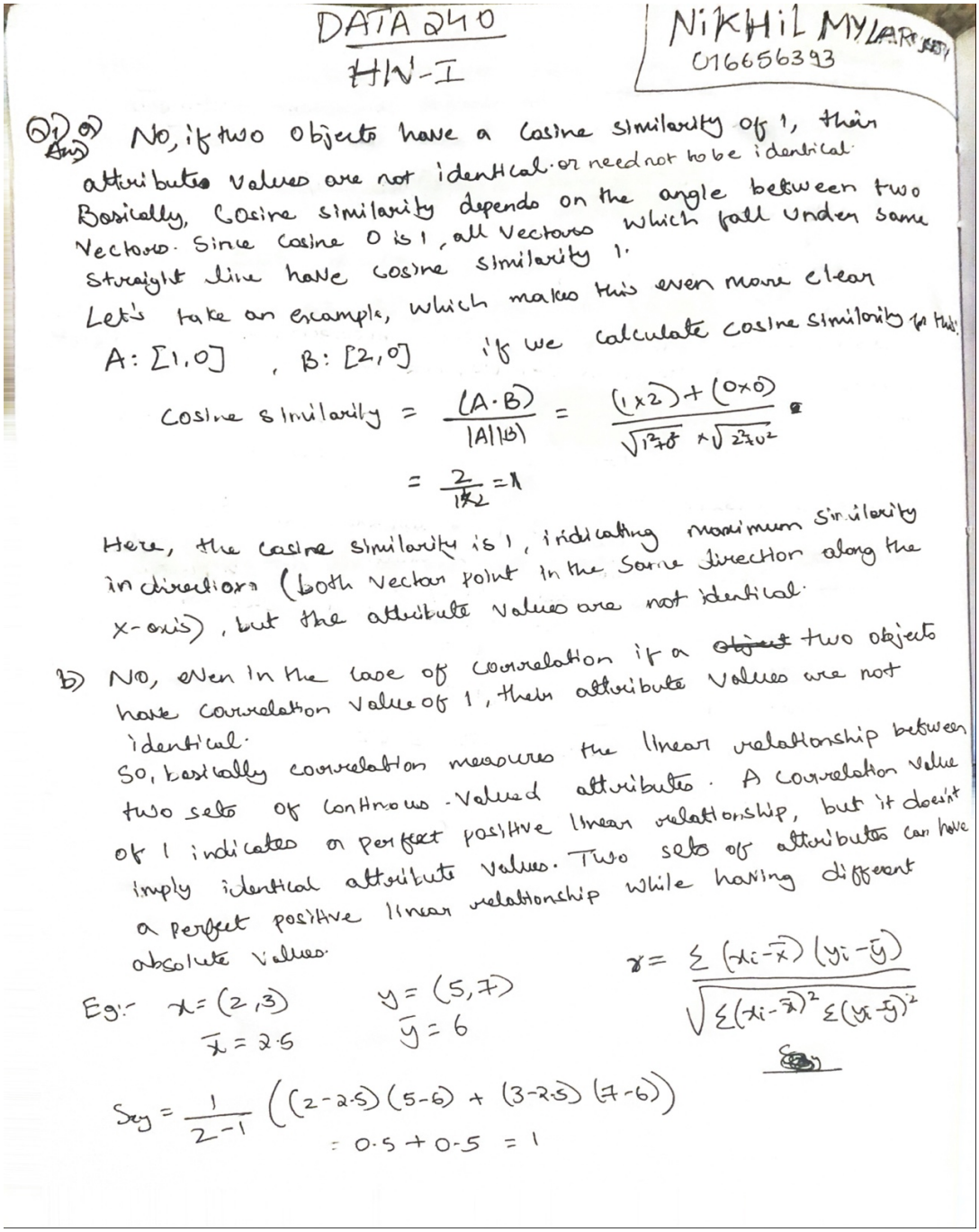
ASSIGNMENT-1(DATA240)

1Q

a) If two objects have a cosine similarity of 1, must their attribute values be identical? Explain.
b) If two objects have a correlation value of 1, must their attribute values be identical? Explain.
c) If two objects have a Euclidean distance of 0, must their attribute values be identical?Explain.



DATA 240
HW-I

NIKHIL MYLARI
016656393

Q1) a) No, if two objects have a cosine similarity of 1, their attributes values are not identical. or need not to be identical.
Basically, Cosine similarity depends on the angle between two Vectors. Since cosine 0 is 1, all vectors which fall under same straight line have cosine similarity 1.
Let's take an example, which makes this even more clear
A: [1,0] , B: [2,0] if we calculate cosine similarity for this

$$\text{Cosine similarity} = \frac{(A \cdot B)}{|A||B|} = \frac{(1\times2)+(0\times0)}{\sqrt{1^2+0^2}\times\sqrt{2^2+0^2}}$$

$$= \frac{2}{1\times2} = 1$$

Here, the cosine similarity is 1, indicating maximum similarity in direction (both vector point in the same direction along the x-axis), but the attribute values are not identical.

b) No, even in the case of correlation if a ~~object~~ two objects have correlation value of 1, their attribute values are not identical.
So, basically correlation measures the linear relationship between two sets of continuous-valued attributes. A correlation value of 1 indicates a perfect positive linear relationship, but it doesn't imply identical attribute values. Two sets of attributes can have a perfect positive linear relationship while having different absolute values.

Eg:- $x = (2,3)$ $y = (5,7)$
$\bar{x} = 2.5$ $\bar{y} = 6$

$$r = \frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2 \sum(y_i-\bar{y})^2}}$$

$$S_{xy} = \frac{1}{2-1}\left((2-2.5)(5-6) + (3-2.5)(7-6)\right)$$
$$= 0.5 + 0.5 = 1$$

$S_{xy} = 1$

$S_x = \sqrt{\frac{1}{2-1}(2-2.5)^2 + (3-2.5)^2} = \sqrt{(-0.5)^2 + (0.5)^2} = 0.5\sqrt{2}$

$S_y = \sqrt{\frac{1}{2-1}\left[(5-6)^2 + (7-6)^2\right]} = \sqrt{1^2 + 1^2} = \sqrt{2}$

$$\frac{S_{xy}}{S_x S_y} = \frac{1}{(0.5\sqrt{2})(\sqrt{2})} = 1$$

In the above example, the correlation between X & Y is 1, indicating a perfect positive linear relationship, but the attribute values themselves are not identical.

c) Yes, if two objectives have a Euclidean distance of 0, their attributes values must be identical.

Basically, the Euclidean distance measures the straight-line distance between two points in a multi-dimensional space. When the Euclidean distance is 0, it indicates that the two points are at the same location in the space, which means their attributes values are the same.

For example:- $x = (1,1)$ $y = (2,2)$

Euclidean dist = $\sqrt{(2-1)^2 + (2-1)^2} = \sqrt{2}$

but it is same like $x = (2,2)$ $y = (2,2)$

$E.D = \sqrt{(2-2)^2 + (2-2)^2} = 0$.

d) Translation

d) Translation:

Cosine similarity is not invariant in the transformation $x \to x+c$ & $y \to y+c$. This is because the cosine similarity will be changed only when a constant is added to it.

As mentioned the formula in the previous question,

$$\text{Cosine}(x,y) = \frac{x \cdot y}{\|x\| \|y\|}$$

where the transformed vectors are $x' = x+c$ & $y' = y+c$

so $$\text{cosine}(x',y') = \frac{(x+c)\cdot(y+c)}{(\|x+c\| \|y+c\|)}$$

$$= \frac{(x\cdot y + cx + cy + c^2)}{(\|x+c\| \|y+c\|)}$$

$$= \frac{(x\cdot y + c(x\cdot 1) + c(y\cdot 1) + c^2}{(\|x+c\| \|y+c\|)}$$

Though the cosine similarity changes based on the value of c this is not invariant under the given transformation.

That is the reason $\text{cosine}(x,y) \neq \text{cosine}(x',y')$.

Regarding the Correlation, the translation of two vectors by a constant value will not be affected as the correlation is the linear measure of the two variables associated & it is invariant to the location of the variables in the number line.

$$\text{Corr}(x,y) = \frac{\text{covariance}(x,y)}{S.D(x) \times S.D(y)} = \frac{S_{xy}}{S_x S_y}$$

i.e $S_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$, $S_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$ & S_y.

Where the transformed vectors are $x' = x+c$ & $y' = y+c$

Sub the transformed in formula above

$$\text{Corr}(x,y) = \frac{\text{covariance}(x',y')}{S.D(x) \cdot S.D(y)}$$

subst $x' = x+c$ & $y = y+c$

$$\text{Covariance}(x,y) = \text{covariance}(x',y')$$

$$S.D(x) = S.D(x+c)$$

$$S.D(y) = S.D(y+c)$$

$$\text{Corr}(x',y') = \frac{\text{covariance}(x+c, y+c)}{SD(x+c)\ S.D(y+c)}$$

$$= \text{corr}(x,y)$$

But by adding constants won't affect the covariance rather it only changes the scales of the covariance. Even S.D is not affected.

So, $\text{Corr}(x,y) = \text{Corr}(x',y')$

Euclidean Distance is also invariant under translation. It measures the straight line distance between points & adding a constant c to each attribute value does not change this distance.

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$$

Sub $x' = x+c$ & $y' = y+c$

$$d(x',y') = \sqrt{\sum_{i=1}^{n}(x_{i+c}-y_{i+c})^2} = \sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$$

Thus E.D $d(x,y) = d(x',y')$.

Scaling:

Scaling :- $x \to cx$ & $y \to cy$, where c is a constant multiplied to each attribute value in x & y.

i) Cosine similarity : It is invariant under scaling. Scaling all attributes by a constant c does not change the angle between vectors.

Eg:- $x = [1,2]$ $y = [3,4]$

$$\frac{1(3)+2(4)}{\sqrt{1^2+2^2}\sqrt{3^2+4^2}} = \frac{11}{\sqrt{5}\sqrt{25}} = 0.9759$$

Now, lets scale x & y by a constant $c=2$

$x' = cx = 2[1,2] = [2,4]$

$y' = cy = 2[3,4] = [6,8]$

$$\frac{2(6)+4(8)}{\sqrt{3^2+4^2}\sqrt{6^2+8^2}} = 0.9759.$$

Correlation in scaling will not change the linear relationship between two vectors as correlation is invariant under scaling transformation.

$$\text{Correlation}(x,y) = \frac{\text{covariance}(x,y)}{\text{std}(x) \cdot \text{SD}(y)}$$

Sub $x' = cx$, $y' = cy$

$$\text{Correlation}(x',y') = \frac{\text{covariance}(x',y')}{\text{S.D}(x')\,\text{S.D}(y')}$$

$$= \frac{\text{covariance}((x,y)}{\text{SD}(cx)\,\text{SD}(cy)}$$

$$= \frac{c^2\,\text{covariance}(x,y)}{c\,\text{SD}(x)\,c\,(\text{SD}y)}$$

So, from this we can see that correlation $(x',y') = \text{Corr}(x,y)$

iii) Euclidean Distance is affected by scaling. Scaling the attributes by a constant c scales the Euclidean distance by |c|.

For example :- $x[1,2]$ $y[3,4]$

Euclidean distance $= \sqrt{(1-3)^2+(2-4)^2} = \sqrt{4+4} = \sqrt{8} = 2.83$.

Now, if we scale both x & y by a constant $c=2$:

$x' = cx = 2[1,2] = [2,4]$ $\quad y' = cy = 2[3,4] = [6,8]$

$\sqrt{(2-6)^2+(4-8)^2} = \sqrt{16+16} = \sqrt{32} = 5.66$.

So, we can see the values got changed in this case! so it is affecting.

Standardization:

Standardization:

Cosine similarity, the standardization would change the distance between two vectors. This is because the vectors will be scaled by unit length & angle between them changed.

Thus the cosine similarity will change as the c values changes, so it is variant under given standardization.

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

Sub $x' = \frac{x-c}{d}$, $y' = \frac{y-c}{d}$

$$\cos(x',y') = \frac{\left(\frac{x-c}{d}\right)\left(\frac{y-c}{d}\right)}{\left\|\frac{x-c}{d}\right\| \left\|\frac{y-c}{d}\right\|}$$

$$= \frac{(x-c)(y-c)/d^2}{\frac{\|x-c\| \|y-c\|}{d^2}}$$

i.e $\cos(x',y') \neq \cos(x,y)$.

(ii) **Correlation** :- Standardization would not change the proximity between the two vectors & therefore due to the linear relationship btwn them it will remain the same.

$$Corr = \frac{Cov(x,y)}{S.D(x)\,S.D(y)}$$

Sub $x' = \frac{x-c}{d}$, $y' = \frac{y-c}{d}$

$$cor(x',y') = \frac{Cov\left(\frac{x-c}{d}, \frac{y-c}{d}\right)}{S.D\left(\frac{x-c}{d}\right) S.D\left(\frac{y-c}{d}\right)}$$

$$= \frac{1/d^2\left(cov(x-c, y-c)\right)}{1/d^2\left(SD(x-c)\cdot SD(y-c)\right)}$$

$$Cor(x',y') = Cor(x,y)$$

So that's the reason correlation is invariant under standardization.

(iii) But Euclidean distance is affected by standardization with constants c & d.

Eg:- Consider two objects, $x = [1,2]$ & $y = [3,4]$. Let's standardize x & y using constants $c=2$ & $d=2$

$$x' = \frac{(x-c)}{d} = \frac{([1,2]-[2,2])}{[2,2]} = [-1/2, 0]$$

$$y' = \frac{(y-c)}{d} = \frac{([3,4]-[2,2])}{[2,2]} = [1/2, 1]$$

Now, lets calculate the E.D between x' & y':

$$\sqrt{\left(\left(-\frac{1}{2}-\frac{1}{2}\right)^2 + (0-1)^2\right)} = \sqrt{0^2 + (-1)^2} = \sqrt{1} = 1$$ & without standardization is 2.83

As, we can see standardization with constants c & d affects the E.D distance by $|d|$, & in this example, it resulted in a E.D of 1.

However, only Correlation remained unchanged under Standardization.

2Q

2) Positivity :- Distance measures satisfies the positivity property.

$$D(X,Y) = \min \{ d(x,y) : x \in X, y \in Y \}$$

This property is satisfied because the Euclidean distance $d(x,y)$ is always non-negative, and taking the minimum of non-negative values will also result in a non-negative value.
Therefore, $D(X,Y) \geq 0$ for any clusters X & Y.

Symmetric property :- Distance ~~also~~ measures also satisfies the symmetric property.

$$D(X,Y) = \min \{ d(x,y) : x \in X, y \in Y \} = \min \{ d(y,x) : x \in X, y \in Y \} = D(Y,X)$$

~~$D(X,Y) = \min \{ d(x,y) \} \; x \in X \; D(Y,X)$~~

so, basically symmetry requires that the distance between clusters X & Y is the same as the distance between clusters Y & X

$(D(X,Y) = D(Y,X))$.

D measures the distance between clusters in terms of the closest two points from each cluster. Since the E.D is always same when we measure distance between X & Y, Y & X because the E.D formula contains square of the difference. So, there is no change in the distance value.

Triangle Inequality :

The triangle inequality requires that the distance between clusters X & Z is always less than or equal to the sum of the distances between X & Y and between clusters Y & Z

$$D(X,Z) \leq D(X,Y) + D(Y,Z)$$

For example of a counter :

Basically the triangle inequality is not guaranteed to hold for this distance measure

$X = (0,0)$ $Y = (1,1)$, $Z = (2,2)$

The distance are :

$D(X,Y) = \min \{ d(0,1) \} = \min \{ 1 \} = 1$

$D(Y,Z) = \min \{ d(1,2) \} = \min \{ 1 \} = 1$

$D(X,Z) = \min \{ d(0,2) \} = \min \{ 2 \} = 2$

$D(X,Z)$ is not less than or equal to $D(X,Y) + D(Y,Z)$, violating the triangle inequality.

So, the distance measure is not guaranteed to satisfy the triangle inequality.

3Q

a) Explain why cosine is not a good measure for clustering the data
b) Explain why correlation is not a good measure for clustering the dataExplain
c) what preprocessing steps and corresponding proximity measure you should use cluster the data.

3) a) Cosine similarity is not a suitable measure for clustering this Census data for several reasons. Cosine similarity is primarily used for text or high-dimensional data, where the focus is on the direction or angles between vectors. In this census data with continuous attributes (which have wide range of values) such as total household income, property value, & no. of bedrooms, the concept of direction or angles between vectors is not meaningful. Cosine similarity may also not handle the varying scales of attributes well, as some attributes (like income) may have much larger values than others (like the no. of bedrooms), potentially leading to biased results. Therefore, cosine similarity may not capture the meaningful relationships between household based on their attributes.

b) Correlation may not be an ideal measure for clustering this Census data because it is primarily used to measure linear relationships between attributes. While its valuable for identifying linear dependencies, it may not capture complex relationships that can exist between attributes in a household, which may be nonlinear in nature. Additionally, correlation assumes that the data follows a normal distribution, which might not be the case for all attributes in the Census data. This can lead to inaccurate clustering results.

c) To cluster the census data effectively, its essential to prepare the data through preprocess steps & choose an appropriate proximity measure. First, we should standardize the data, which means giving all attributes a common scale with zero mean & unit variance. This prevents attributes with differing scales from dominating the clustering process. Additionally, identifying & handling outliers is crucial to ensure they don't distort the clustering results; methods like z-score or IQR can be helpful. Lastly, even normalization may be necessary, particularly if attributes have varying ranges, as it scales attributes to a common range, often [0,1].

In terms of proximity measures, given that the census data comprises continuous attributes, the Euclidean distance is a fitting choice. The Euclidean distance calculates the direct distance between data points in a multi-dimensional space, considering both attribute magnitudes & differences. This makes it well-suited for uncovering clusters based on attribute values. By following these preprocessing steps & adopting the E.D as the proximity measures, we can cluster the census data effectively, capturing attribute variations & relationships meaningfully.

So, basically this is how it goes :-

$$D(A,B) = \sqrt{\sum_{i=1}^{n}(A_i - B_i)^2}$$

Where $D(A,B)$ is the E.D between points A & B

n is the no. of dimensions

A_i & B_i are the standardized values of attribute i for points A & B.

4Q

i) **positivity** : It states that distance is always positive i.e non negative

(i) $d(x,y) \geq 0$ for all x & y

(ii) $d(x,y) = 0$ if & only if $x = y$

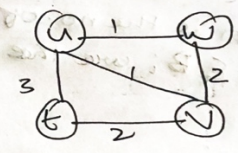
Given that $d(u,v) = 0$ if $u = v$, since the distance can never be negative, $d(u,v) \geq 0$ for all $u \& v$.

Both the positivity cases mentioned above are satisfied. (i & ii)

Symmetry : Symmetry requires that the distance between $u \& v$ is the same as the distance between $v \& u$ i.e $d(x,y) = d(y,x)$ for all $x \& y$.

i.e $d(u,v) = d(v,u)$

$d(u,v)$	$= d(v,u)$
$u \to v$	$v \to u$
$u \to w \to v$	$v \to w \to u$
$u \to t \to v$	$v \to t \to u$



We can reach u to v in three different ways & also v to u in three different ways.

Therefore, the symmetric property is also satisfied.

Triangle Inequality: This requires that the distance between U&V is always less than or equal to the sum of the distance between u & an intermediate node w, & between w & V

$d(u,v) \leq d(u,w) + d(w,v)$

So lets take example:-

$1 \leq 3+2 \rightarrow$ so this is satisfied

$d(u,t) \leq d(u,v) + d(v,t)$

$7 \leq 1+5 \rightarrow$ Where this is not satisfied

Since all properties are not getting satisfied $d(u,v)$ is not a metric.

5Q

5) So, the missing values can be filled with the column averages:

$$\frac{-0.2326 - 0.0847 + 0.1275 + 0.3724 + 0.4975 + 0.6926 + 0.7953 + 0.8229 + 0.8497 + 1.0592 + 1.5028}{11}$$

$$= \frac{6.4006}{11} = 0.5818$$

11'y the average of second column after removing the missing value is $\frac{5.4586}{11} = 0.49623$

So, $A_{4,1}$ & $A_{11,2}$ are 0.5818 & 0.49628 respectively.

Let's put that in the RMSE equation

$$RMSE = \sqrt{\frac{(0.1329 - 0.5818)^2 + (0.76 - 0.49623)^2}{2}}$$

Where $A_{4,1} = 0.1329$, $A_{11,2} = 0.76$

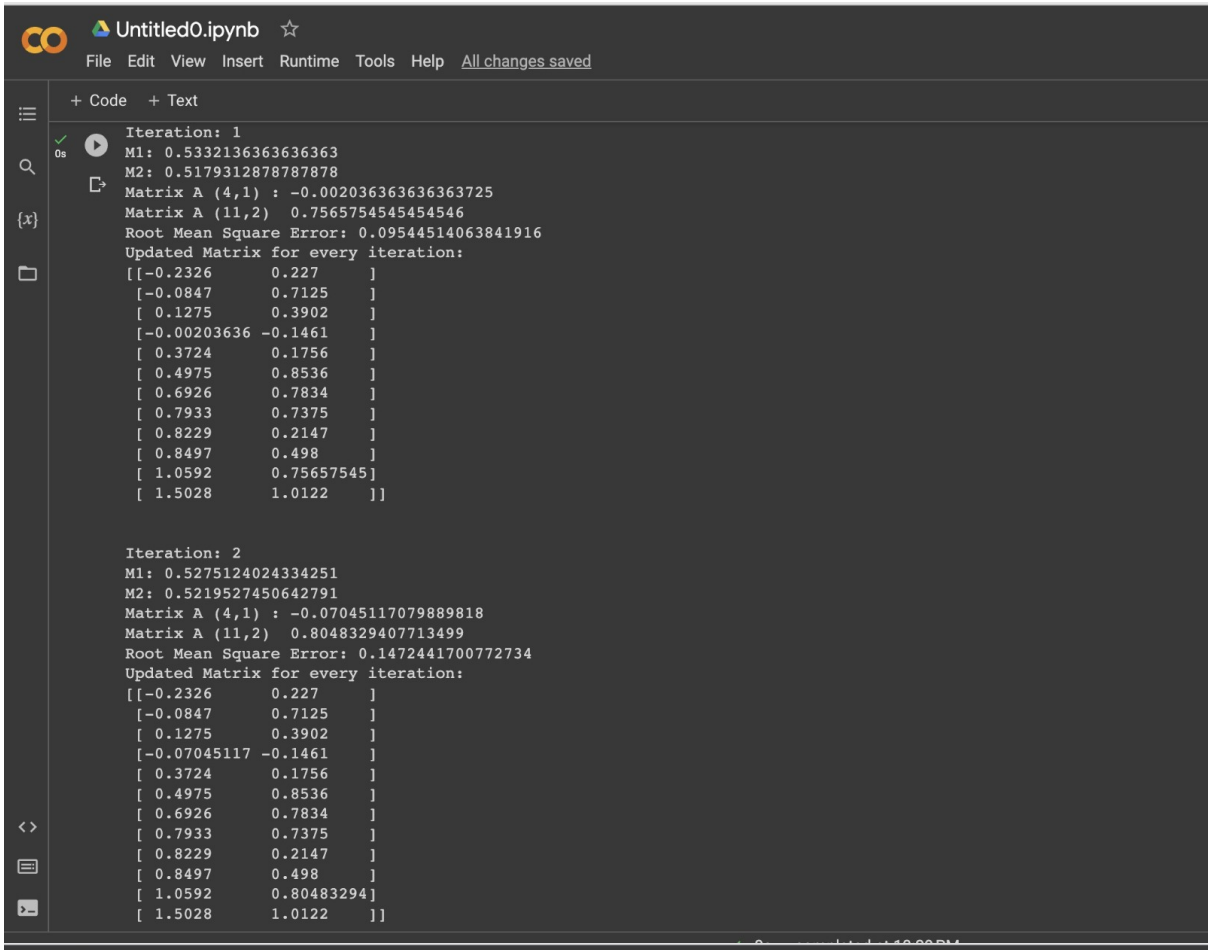
$\bar{A}_{4,1} = 0.5818$, $\bar{A}_{11,2} = 0.49623$

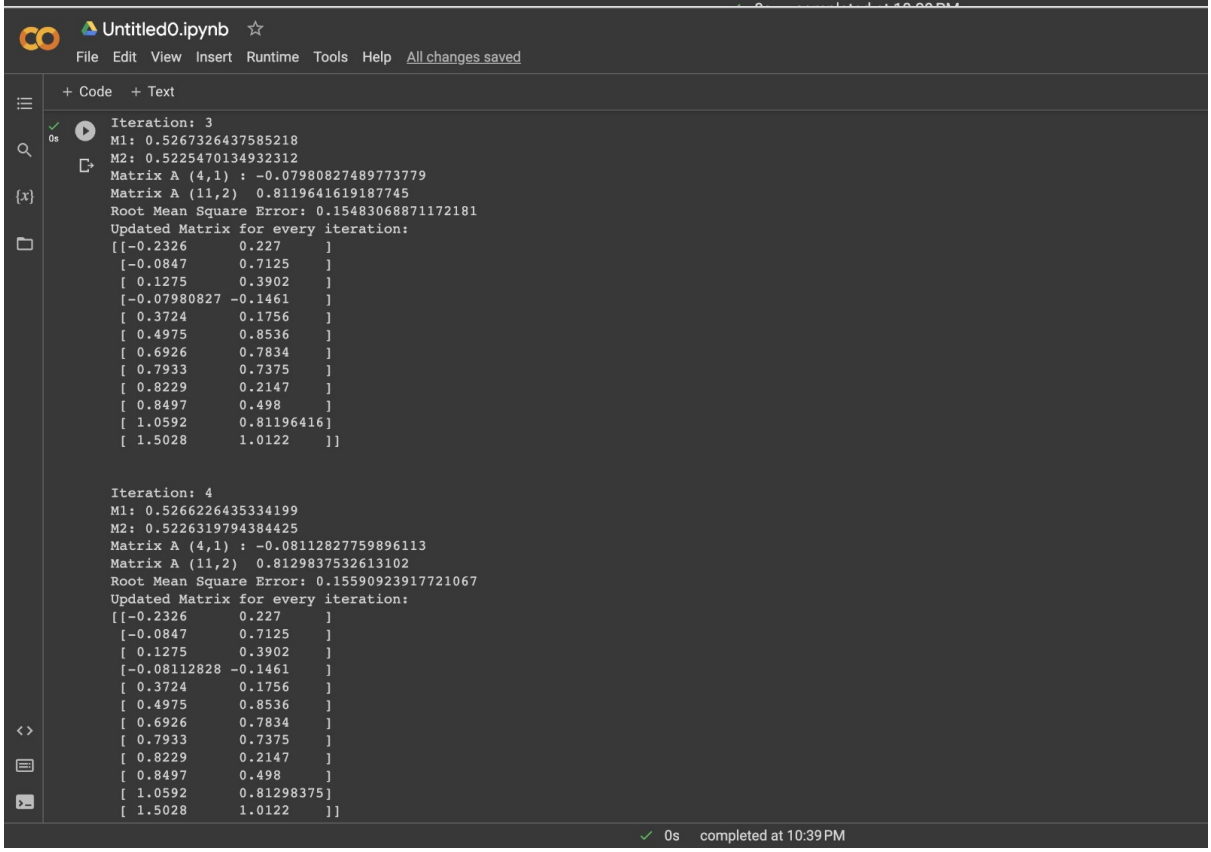
$$= \sqrt{\frac{(-0.4489)^2 + (0.264)^2}{2}} = \sqrt{\frac{0.2015 + 0.0696}{2}} = \sqrt{\frac{0.271196}{2}}$$

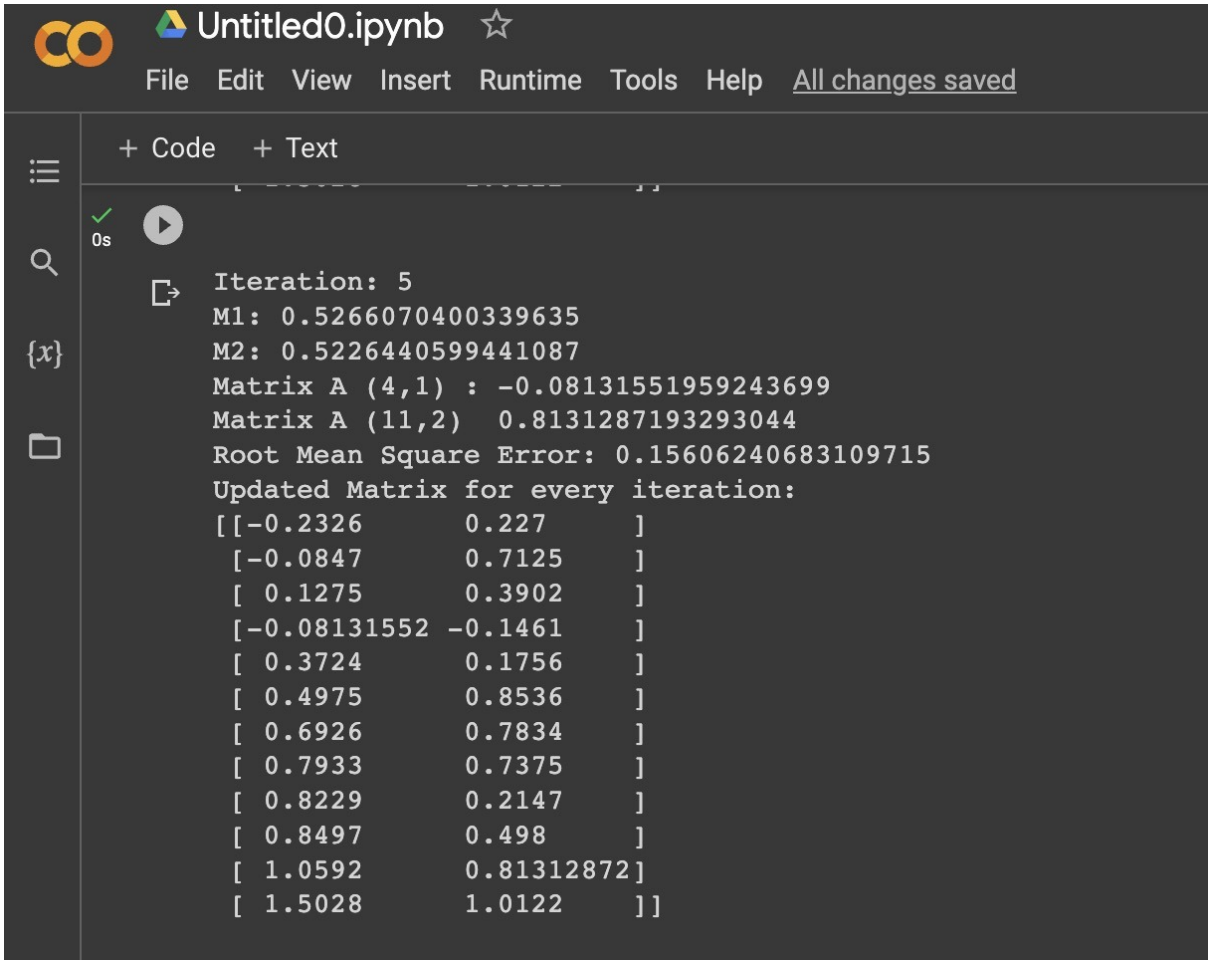
$$= 0.3683$$

a)

b)


Untitled0.ipynb
File Edit View Insert Runtime Tools Help All changes saved
+ Code + Text
Iteration: 1
M1: 0.5332136363636363
M2: 0.5179312878787878
Matrix A (4,1) : -0.002036363636363725
Matrix A (11,2) 0.7565754545454546
Root Mean Square Error: 0.09544514063841916
Updated Matrix for every iteration:
[[-0.2326 0.227]
[-0.0847 0.7125]
[0.1275 0.3902]
[-0.00203636 -0.1461]
[0.3724 0.1756]
[0.4975 0.8536]
[0.6926 0.7834]
[0.7933 0.7375]
[0.8229 0.2147]
[0.8497 0.498]
[1.0592 0.75657545]
[1.5028 1.0122]]
Iteration: 2
M1: 0.5275124024334251
M2: 0.5219527450642791
Matrix A (4,1) : -0.07045117079889818
Matrix A (11,2) 0.8048329407713499
Root Mean Square Error: 0.1472441700772734
Updated Matrix for every iteration:
[[-0.2326 0.227]
[-0.0847 0.7125]
[0.1275 0.3902]
[-0.07045117 -0.1461]
[0.3724 0.1756]
[0.4975 0.8536]
[0.6926 0.7834]
[0.7933 0.7375]
[0.8229 0.2147]
[0.8497 0.498]
[1.0592 0.80483294]
[1.5028 1.0122]]


Untitled0.ipynb
File Edit View Insert Runtime Tools Help All changes saved
+ Code + Text
Iteration: 3
M1: 0.5267326437585218
M2: 0.5225470134932312
Matrix A (4,1) : -0.07980827489773779
Matrix A (11,2) 0.8119641619187745
Root Mean Square Error: 0.15483068871172181
Updated Matrix for every iteration:
[[-0.2326 0.227]
[-0.0847 0.7125]
[0.1275 0.3902]
[-0.07980827 -0.1461]
[0.3724 0.1756]
[0.4975 0.8536]
[0.6926 0.7834]
[0.7933 0.7375]
[0.8229 0.2147]
[0.8497 0.498]
[1.0592 0.81196416]
[1.5028 1.0122]]
Iteration: 4
M1: 0.5266226435334199
M2: 0.5226319794384425
Matrix A (4,1) : -0.08112827759896113
Matrix A (11,2) 0.8129837532613102
Root Mean Square Error: 0.15590923917721067
Updated Matrix for every iteration:
[[-0.2326 0.227]
[-0.0847 0.7125]
[0.1275 0.3902]
[-0.08112828 -0.1461]
[0.3724 0.1756]
[0.4975 0.8536]
[0.6926 0.7834]
[0.7933 0.7375]
[0.8229 0.2147]
[0.8497 0.498]
[1.0592 0.81298375]
[1.5028 1.0122]]
0s completed at 10:39 PM



The ultimate table displaying the imputed values and corresponding RMSE for each iteration.