G15_HW1

Team Members:

Darpan Dodiya - dpdodiya
Shantanu Sharma - ssharm34
Shrijeet Joshi - sjoshi22

Solutions

Solution 1

A.

Sno	Attribute	Type	Assumption	Classification
1	Blood Group (A,B,AB,O)	Nominal		Discrete
2	Ticket number for raffle draws	Nominal	Ticket # is randomly generated and has no seq	Discrete
3	Brightness as measured by a light meter	Ratio		Continuous
4	Grade in terms of Pass or Fail	Nominal		Binary
5	Time zones (EST, PST, CST)	Interval		Discrete
6	Income earned in a month	Ratio		Continuous
7	Vehicle license plate number	Nominal		Discrete
8	Distance from the center of campus.	Ratio		Continuous
9	Dorm room number	Ordinal	Numbers are given in order eg G01, G02	Discrete
10	Kelvin temperature	Ratio		Continuous

B

Sno	Attribute	Type	Operations	Assumptions
1	Make	Nominal	Mode	All types are independent and no 1 type is better than other
2	Fuel Type	Nominal	Mode	
3	Doors	Ratio	Mean,Mode,Median,Pearson correlation, standard deviation, Binary Discretization, z score normalization	
4	Height	Ratio	Mean,Mode,Median,Pearson correlation, standard deviation, Binary Discretization, z score normalization	
5	Cylinders	Ratio	Mean,Mode,Median,Pearson correlation, standard deviation, Binary Discretization, z score normalization	
6	Price	Ratio	Mean,Mode,Median,Pearson correlation, standard deviation, Binary Discretization, z score normalization	

C.

Test Score can be considered Ordinal attribute if we look to rank students on the basis of test scores and we do not know how the test was graded. This is similar to assigning A,B+,B grades etc. We do not know the exact raw score of the student based on his score from 0 -5.
E.g. We wish to find if student A performed better than B if his/her score is more than the other.

Test Scores can be considered Ratio Attribute,when we wish to scale the student marks to some standardized scale eg 100.

Solution 2

(A.)

ID	Patient	Treatment A(SBP)	Treatment B(SBP)
1	Patient 1	160	300
2	Patient 2	120	100
3	Patient 3	130	NA
4	Patient 4	NA	130
5	Patient 5	120	110
6	Patient 6	NA	100
7	Patient 7	240	120
8	Patient 8	140	90

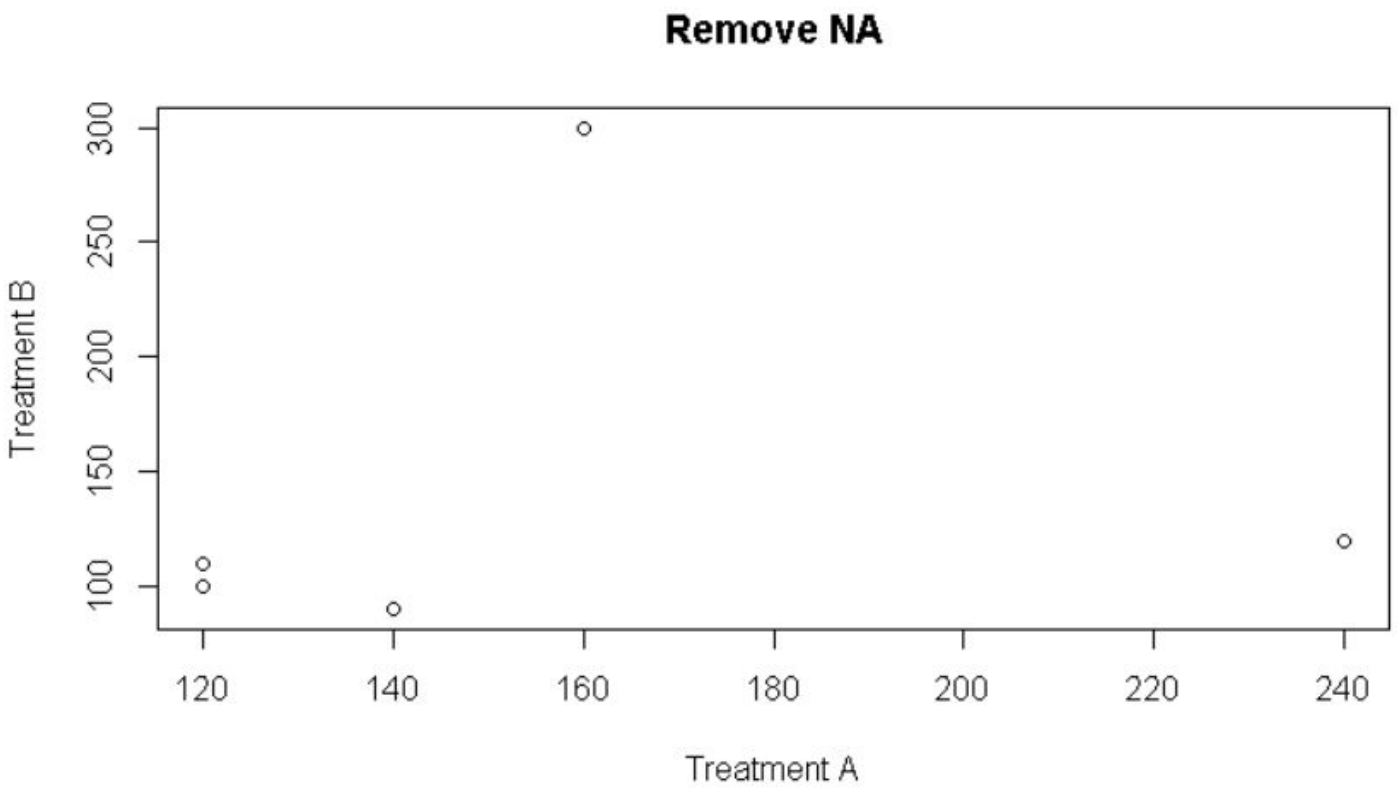
(B.)
Strategy 1: Remove patients with NA values : This strategy leads to loss of data and is okay is # removed rows far less than total sample size. In our case if we decide to remove rows with NA, we are removing 3 of 8 rows i.e. 37% data.

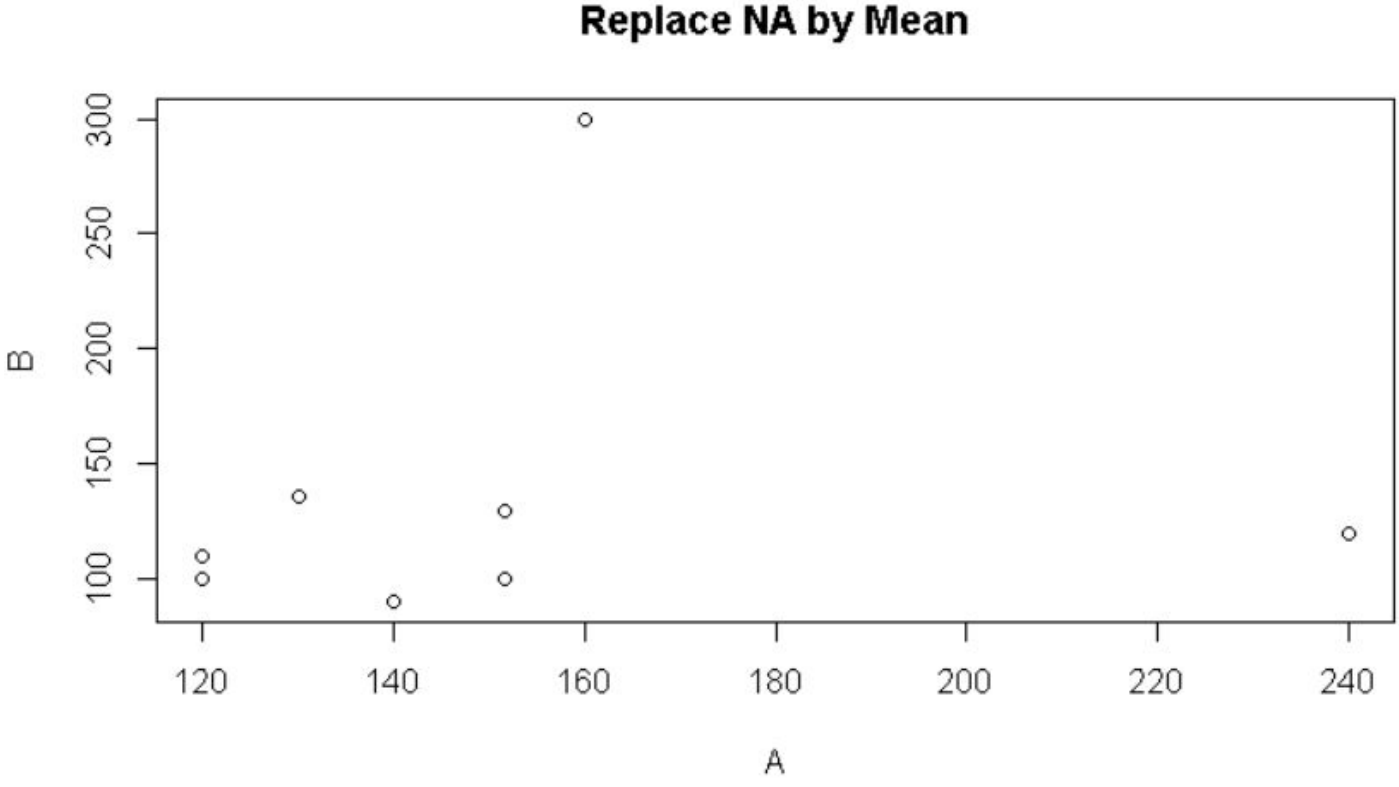
Strategy 2: Replace NA with mean of other patients. This is an easy approach to handle missing data and we do not incur any dataset loss but reduces the variability of data and addition of bias.

The better approach in this case is Strategy 2 as the dataset provided to us is already quite small, and removal of rows will make it smaller even further which would be catastrophic.

(C.)
Noise is deviation from the true value, it may have values close to the true value, where as outlier is something that is much different than the other values in the dataset. The vast majority of time outliers are noise but sometimes a data point that is true signal can be an outlier

1 For this dataset .
240, and 300 are more likely to be outliers, a visual examination of the plots below confirm the case.


Remove NA
Treatment B
100
150
200
250
300
120
140
160
180
200
220
240
Treatment A


Replace NA by Mean
B
100
150
200
250
300
120
140
160
180
200
220
240
A

Strategies to deal with outliers :

1. Trimming : Dataset excludes the outlier
2. Value Transformation : Run square root or log transformation to the outlier values so that their impact is reduced.
3. Outliers might also need to be examined before being rejected as they could suggest anomaly. For eg SBP 240/300 might be a reading of a patient with severe medical condition and might require immediate medical condition.

C. The +/- 3 variations introduce noise to the dataset. As we can see a variation of +/- 3 will generate new values close to the true value (150).
Strategies to deal with this :

1. Taking the average of several readings from the instrument.
2. Replace the instrument.

Solution 3

Q3] Sampling (7 points)
Answer
A] State the sampling method used in the following scenarios and give a reason for your answer. Choose from the following options: simple random sample with replacement, simple random sample without replacement, stratified sampling, progressive/adaptive sampling.

1. Stratified Sampling

Stratified sampling is a good approach as we are able to group professors into homogenous strata according to the given groups. This is a good approach as each group will have a similar salary. Further, simple random sampling in each strata can be done which will overall give a good estimate of the average salary.

2. Simple random sample with replacement

Sample of {2,2,2,6,8} was collected from population {2,2,4,4,6,6,8}. We can see that 2 was repeated more than the number of times it is seen in the population, this points to with replacement sampling.

3. Progressive/adaptive sampling

We require an end result of 90% accuracy, however we do not know how many samples will be required to reach this outcome. Hence, sampling is continued progressively until such an outcome is reached.

B] The U.S. Congress is made up of 2 chambers: 1) a Senate of 100 members, with 2 members from each state, and 2) a House of Representatives of 435 members, with members from each state proportional to that state's population. For example, Alaska has 2 Senators and 1 House representative, while Florida has 2 Senators and 27 House representatives. Both the Senate and

the House are conducting surveys of their constituents, which they want to reflect the makeup of each chamber. You suggest that they use stratified sampling for this survey, sending surveys to a certain number of people from each state. Each survey will be sent to 1000 participants.

1. When subpopulations within a population vary significantly, it is better to sample each subpopulation independently. In the above use-case, the populations in the states from which the members are elected vary significantly. Also, such subpopulations of the states are independent and collectively exhaustive as a person can belong to a single state only as per some official documentation such as driving license etc.
If we conduct simple random sampling, states with less population might be left out and this will induce sampling error.
Hence, stratified sampling is appropriate here. This will help to increase the precision by reducing sampling error. It produces a weighted mean that has less variability than an overall arithmetic mean.

2. Senate has 100 members, 2 from each state. Hence, senate members are allocated equally across all states without considering the population of each state. To reflect this makeup, I will recommend sending the survey equally across each state, sending (2/100)*1000 = 20 surveys to each state including Alaska. This is an example of disproportionate stratified sampling.

3. House of Representatives has 435 members which are allocated according to the population of every state. Hence, a state of higher population will have a greater number of representatives than a state with a lower population. To reflect this makeup, I would recommend sending (27/435)*1000 = 62 surveys to Florida. This is an example of proportionate stratified sampling.

4. “House” approach to stratified sampling makes sure that the number of samples in a group are proportional to the group size. This reduces sampling error as no group is sampled too heavily or lightly to create bias. This is advantageous if the population within each strata has heterogeneous values. Then such an approach will provide better precision.
“Senate” approach to stratified sampling is advantageous if the population within each strata has homogeneous or near homogeneous values. Then such an approach will provide better precision.

Solution 4

Q4] Dimensionality Reduction (12 points)
Answer

A] In figure 1, the most reasonable number of components to retain is PC1 as all other components have a gentle slope compared to PC1 which has a steep slope and has eigenvalue of 1.3 which is > 1 by rule of thumb. All other components have eigenvalues <= 1.0. Majority of variance is explained by PC1.

B] In figure 1, according to the first principal component, Petal.Width and Petal.Length are the features that explain the most variation.

C] In figure 2, the most reasonable number of components to retain is PC1 and PC2 as all other components have a gentle slope compared to PC1 and PC2 which have a steep slope and explain the majority of the variation in the data.

D] In figure 2, according to the first principal component, Sepal.Length, Sepal.Width, Petal.Length, Petal.Width are the features that explain the most variation.

E] PCA1 is calculated with the original dataset i.e without z score normalization. Hence, features with higher variance will be given a higher weight in PCA analysis. This can cause some other attributes which are equally important but have lower variance be given lower priority.
However, after z score normalization on the dataset i.e transforming the data to mean 0 standard deviation 1 all attributes are equally important and this makes sure we do not lose out on any important attributes.
This can be seen in the case of Sepal.Length attribute in PCA1 vs PCA2.

Hence, due the the above factors we should use PCA2.

F] Based on the results of PCA1 and PCA2, we would like to select all the features. As seen in PCA2 all the features contribute similarly to the variance in the data. Hence, we should select and retain all the features.

Solution 5 Discretization

5 (a)
4 equal width discretization on TEMPERATURE attribute.

The minimum of all TEMPERATURE values is: 50.0
The maximum of all TEMPERATURE values is: 95.0
Difference between maximum and minimum is (this is also the range of data): 45.0

Since we need 4 equi-width intervals, we can divide the range into 4 equal sections: 45.0 / 4 = 11.25

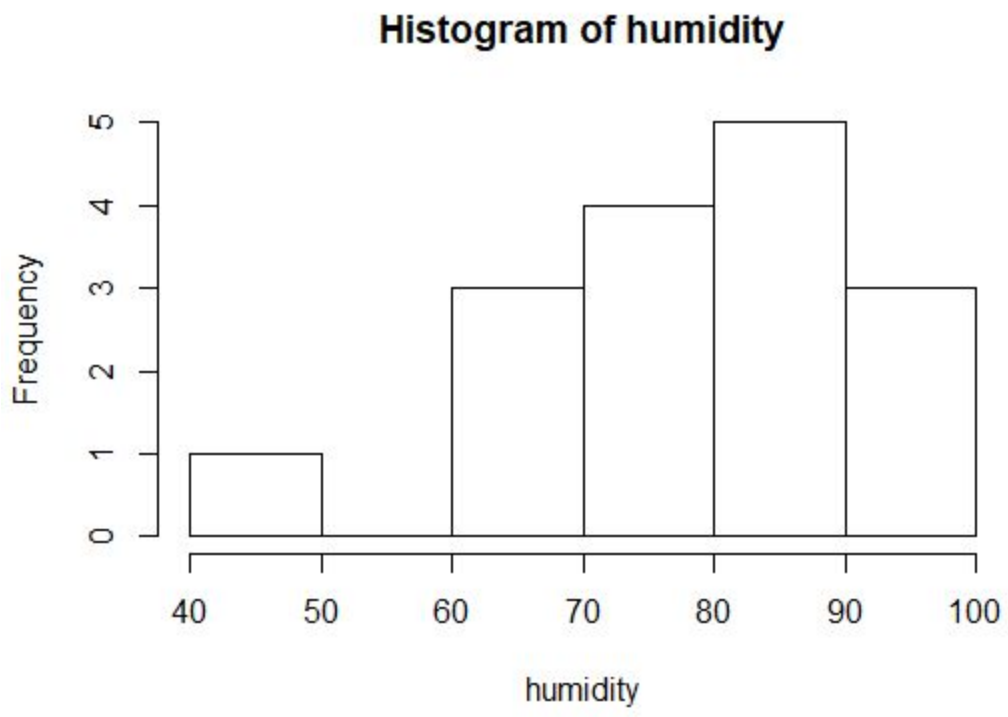
Thus, starting from 50, we can have 4 intervals. The TEMPERATURE values can be binned accordingly. The result is shown below:

Interval	**Data**
[50.0 - 61.25]	50.0
[62.0 - 72.5]	70.0, 68.0, 65.0, 64.0, 72.0, 69.0, 71.0
[72.5 - 83.75]	80.0, 83.0, 75.0, 75.0, 81.0, 73.0
[83.75 - 95.0]	95.0, 85.0

5 (b)
4 equal depth discretization on HUMIDITY attribute.

To divide the data in the 4 groups, the best approach is to look at the histogram of the data.

The histogram of the given data looks as follows.



As it can be seen, the data is sparse till 70.0 and then gets dense. After trying out different intervals, below set would put the data into equi-frequency bins.

Interval	**Data**
[45.00 - 70.75]	70.0, 65.0, 70.0, 45.0
[71.75 - 82.50]	80.0, 80.0, 75.0, 71.0
[82.50 - 89.25]	85.0, 86.0, 89.0, 85.0
[89.25 - 96.00]	90.0, 96.0, 95.0, 91.0

5 (c)
Mean and standard deviation approach

Here the given,
mean = 80
sd = 13

The asked interval can be simplified to the following equation,
interval = mean + k * sd,

where k is an integer.

For valid values of k, the discretization is as follows:

k	**Interval (mean + sd * (k-1), mean + sd *k)**	**Data**
-3	<41.0	None
-2	[41.0 - 54.0)	45.0
-1	[54.0 - 67.0)	65.0
0	[67.0 - 80.0)	70.0, 70.0, 71.0, 75.0
1	[80.0 - 93.0)	85.0, 90.0, 86.0, 80.0, 80.0, 89.0, 91.0, 85.0
2	[93.0 - 103.0)	96.0, 95.0
> 3	> 103.0	None

Note for 5a, 5b, 5c:

Any value that is falling on intervals points is binned as follows:
If the interval is
[a - b) e.g. [10-15)
and there are two values to be binned, x and y, where x=a and y=b then
x will be put in [a - b) interval
y will be put in [b - …) interval

5 (d)
Comparison

	5(a) Equal Width	**5(b) Equal Frequency**	**5(c) Mean and SD**
Advantages	- Simple to implement - There could be only one possible set of intervals for data. (Given that number of bins is fixed). Consistent behavior.	- Divides data into equal distribution - Can handle skewed data distribution effectively (e.g. data with outliers). Not sensitive to outliers.	- Simple to implement - Given mean and sd of the data, there could be just one possible set of intervals. Consistent behavior. - As opposed to equal width method, the number of bins doesn't need to be specified.
Disadvantages	- In case of uneven distribution, some bins can contain much more data than others. - Susceptible to outliers. An outlier can drastically skew the range and thus the bins too.	- Not simple to implement. Requires data exploration and some trial and error. - If there are identical values in the data, it may not be possible to divide data into equal parts. (e.g. data like 10, 10, 10, 13) - There could be multiple possible intervals for the same data	- Same as the problem with equal width distribution, in case of uneven distribution, some bins can contain much more data than others. - Outliers cannot be binned properly. - Most data will be binned around the mean.
Use Case	When the data doesn't contain outliers and a simple discretization method is required.	Useful when the data has outliers. Also useful to divide data into equal partitions.	When the data doesn't contain outliers and a simple discretization method is required. Useful when the number of required bins aren't specified.

Solution 6 Distance Metrics

6 (a)

i) Euclidean distance

- Positive definiteness: Yes
The Euclidean distance formula is dist(x, y) = √(x - y)²
Here, the difference between two points is squared first, thus assuring the resultant distance after taking root is always positive.
dist(x,y) >= 0

- Symmetry: Yes
Likewise the previous explanation, the difference is always squared first.
So, (x - y)² or (y - x)² both will yield the same result.

- Triangle inequality: Yes
Euclidean distance method in essence is based upon the premise of Pythagorean theorem. And thus it conforms to the triangle inequality property.

ii) Manhattan distance

- Positive definiteness: Yes
The Manhattan distance can be given by formula,

$$d_1(\mathbf{p}, \mathbf{q}) = \|\mathbf{p} - \mathbf{q}\|_1 = \sum_{i=1}^{n} |p_i - q_i|,$$

Reference

As it can be verified from the formula, an absolute value of the difference is taken into consideration so the property of positive definiteness holds true.

- Symmetry: Yes
Since an absolute measure between two values is taken into account, |p-q| or |q-p| both will result in the same output. This property of symmetry is true.

- Triangle inequality: Yes
If the property is strict one, i.e. d(p, r) > d(p,q) + d(q,r) then this wouldn't satisfy.

E.g. For points, A(2,2), B(1,5), C(5,2),
|A-B| = 4
|B-C| = 7

|A-C| = 3

Thus, |B-C| = |A-B| + |A-C| which violates the strict “greater than” triangle inequality.

iii) Divergence function

- Positive definiteness: No

The distance is defined as
$d(A, B) = 1 - |A \cap B|/|A|$.

Set A = {3}
Set B = {3, 5}

Here d(A, B) = 0 but Set A and Set B are not the same. This violates one property of positive definiteness.

- Symmetry: No

Consider two sets,
Set A = {1, 2, 3, 4}
Set B = {2, 3, 7}

d(A,B) = 1 - ½ = 0.5
d(B,A) = 1 - ⅔ = 0.33

Clearly, d(A,B) ≠ d(B,A)

- Triangle inequality: Yes

Trying out various possible combination of sets, this equality holds true.

e.g.
A = {1, 2}, B = {2, 3}, C= {0, 0} has this equality as true.

A = {1, 2}, B = {2, 3}, C= {1, 3} has this equality as true.

A = {1, 1}, B = {1, 1}, C= {1, 2} has this equality as true.

A = {1, 2}, B = {3, 4}, C= {5, 6} has this equality as true.

iv) Cosine distance

- Positive definiteness: No

Take two points,

A(1,1) and B(4,4)

Here since the angle between both A and B is the same, the distance will result in 0.

But as it can be seen, both are different points. d(A, B) = 0 but they aren't the same. Hence, cosine distance doesn't hold positive definiteness.

- Symmetry: Yes
The function of cosine similarity is defined as;
A.B / (||A|| ||B||)

Here both of the operations in the denominator and numerator are commutative and will produce the same measure of similarity if they are interchanged.

Angle between the two vectors will remain the same even if they were interchanged, so the measure of cosine distance is commutative.

- Triangle inequality: No
Take
A = (1, 0)
B = (√2/2,√2/2)
C = (0, 1)

Cosine-similarity (A, B) = √2/2
Cosine-similarity (B, C) = √2/2
Cosine-similarity (A, C) = 0

To prove that triangle inequality doesn't hold, we need to prove
Cosine-distance (A, C) > Cosine-distance(A, B) + Cosine-distance(B, C)

1 - 0 > 1 - √2/2 + 1 - √2/2
0 > 1 - √2

Hence, proved.

b) 1-N-N Classifier

i) Triangle inequality property

ii)
The item to classify is, y.
The training dataset given is X [x_1, x_2, x_3...x_n]

Also,
d(x,y) is very expensive to calculate so it should be avoided.
d(x_i, x_j) is relatively cheap to calculate so it can be used.

The triangle inequality theorem could be stated as,

d(x_i, y) <= d(x_i, x_j) + d(x_j, y)
d(x_j, y) >= d(x_i, y) - d(x_i, x_j)

E.g. d(x_j, y) = 9 would mean that the distance is at least 9.

Here,
d(x_i, y) will be expensive to calculate. Calculate it once.
d(x_i, x_j) can be pre-calculated.

Then calculate (estimation) d(x_j, y) for all possible values of x_j. Note that this is an inexpensive operation because d(x_i, x_j) is already available and we are calculating d(x_i, y) only once.

Now sort all estimated(x_j, y) distances in ascending order. Remember that these distances are estimated distances and have the relationship of ">=". Iterate through the estimated distances, find the actual value of d(y, x_j). Keep counter of minimum distance so far and stop when the remaining estimated values cannot give any better minimum distance. Skip all x_j values after this.

Return x_j, this will be the nearest point to the y.

An example will better illustrate this: (true distance metric here is manhattan and 2d plane is assumed)

x_1 = (10, 0)
x_2 = (0, 4)
x_3 = (7, 0)
y = (0, 0)

i) Pre-calculate distances
d(x_1, x_2) = 14
d(x_2, x_3) = 11
d(x_1, x_3) = 3

ii) Calculate d(x_1, y) = 10

iii) Estimate for all X elements

$d(x_1, y) >= 10 - 0 = 10$
$d(x_2, y) >= 10 - 14 = - 4$
$d(x_3, y) >= 10 - 3 = 7$

iv) Sort [x_2: -4, x_3:7, x_1:10]

v) Iterate over the sorted list to find “real” distances. For the x_2,

$d(y, x_2) = 5$ (note that this is an expensive operation to calculate but is needed because earlier we “estimated” the distance.)

vi) Compare this distance with the list. If there are no other distances that are less than this, stop and return x_2. No other expensive comparisons are needed as they be always higher.

iii)
Best case:
Yes, this approach does reduce the number of comparisons.
As outlined in the example above, we just need 2 expensive comparisons.

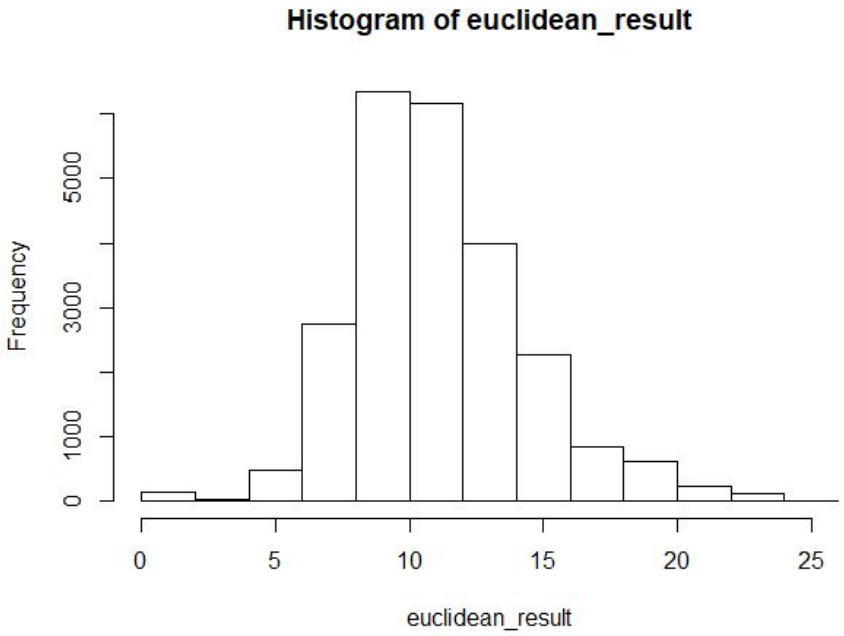
Worst case:
This approach cannot work good in worst case and we’d need |X| comparisons in case we don’t find any element with actual distance less than estimated.

Solution 7

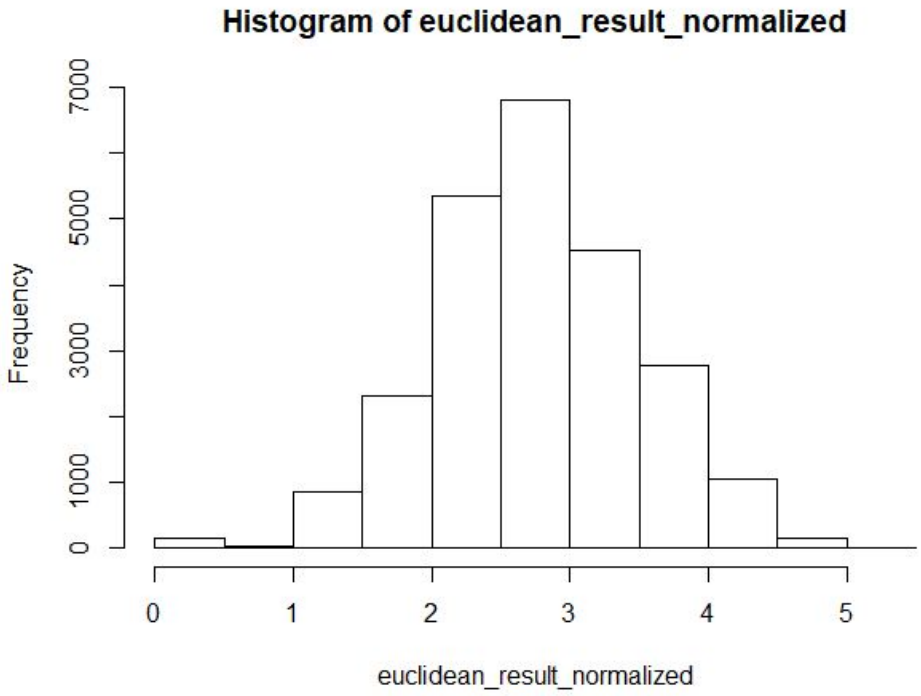
Part 2: Yes, there's a difference between two distance matrices.

a) Using histogram data

Histogram without normalization



Histogram with normalization

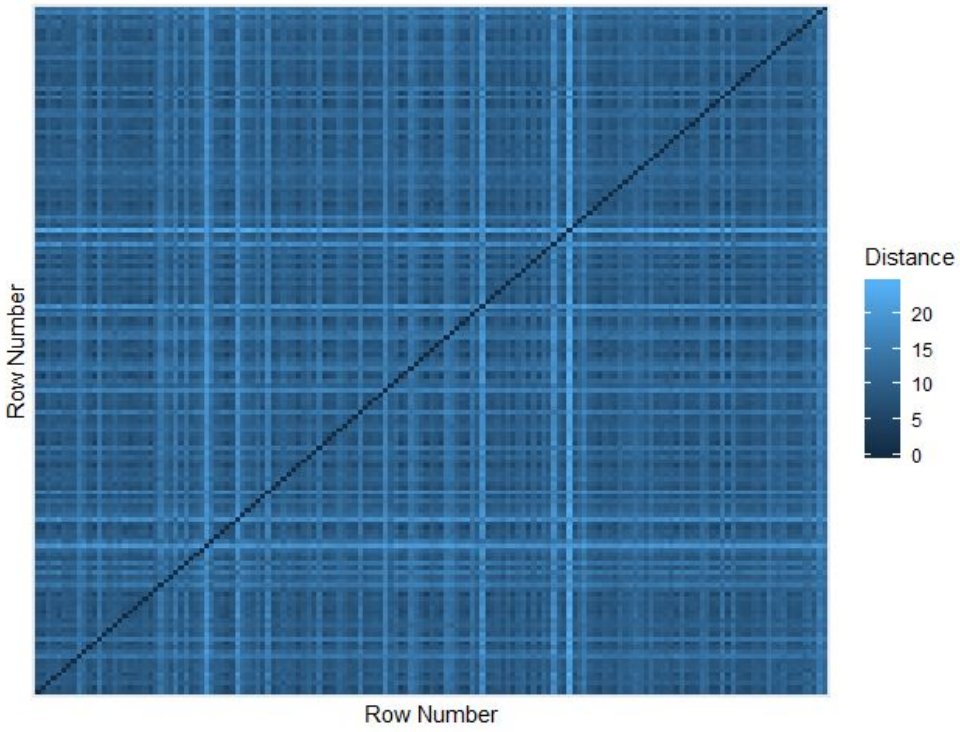


As it can be seen in the above two histograms, histogram of data before normalization is more spread out. (12 bars in 1st chart vs 10 bars in 2nd chart.)

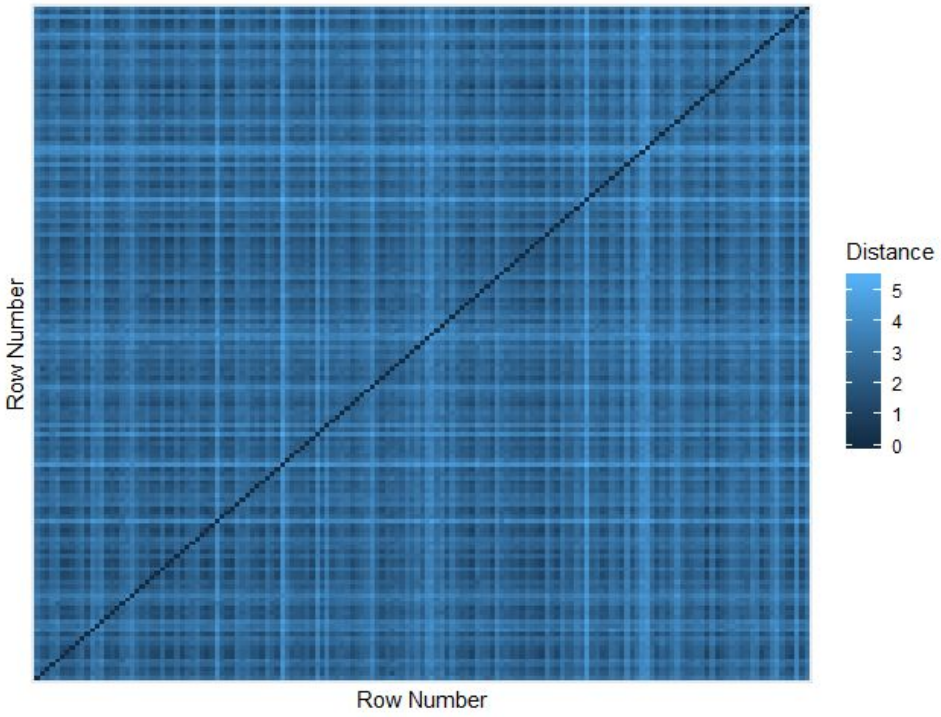
This happened because during normalization, we confined the range of distances to [0, 1] thus reducing spread of data.

b) Using distance visualization

Visualization without normalization



Visualization with normalization



On the first look, there doesn't seem striking difference between the two graphs.

One similarity is the dark blue line connecting left bottom corner to the right upper corner. This line is the distance between the same lines. Because we are using Euclidean distance, the distance has to be 0 for the same vectors. This explains the existence of the straight line. The line will have no effect if the data is normalized or it isn't.

Also due to normalization, light blue lines in normalized graph are more distributed than the data which isn't normalized.

c) Using Min and Max

	Without Normalization	Normalized Data
Min	0	0
Max	24.16609	5.385165

The minimum value remains the same for both versions of data whereas in case of normalized data, the max value reduces from 24.16 to just 5.38.

This can be explained by the normalization operation on the matrix data which limited the data range from 0 to 1.

d) Using Standard Deviation

SD of data before normalization: 3.31837
SD of data before normalization: 0.7477278

As explained in previous points, we have adjusted the scale of the data during normalization process. (reduced the scale) Thus after normalization, the deviation of data too was reduced.