1.

a)

Blood group	Nominal	Discreet	AB, A, B, O, A-
Ticket number for raffle draws	Nominal	Discreet	280740, 192513
Brightness as measured by a light meter	Ratio	Continuous	0.0001,0,400,10000
Grade in terms of Pass or Fail	Nominal	Binary	Pass, Fail
Time zones (EST, PST, CST)	Ordinal	Discreet	PST<CST<EST
Income earned in a month	Ratio	Continuous	0, $1000, $2000
Vehicle license plate number	Nominal	Discreet	
Distance from the center of campus	Ratio	Continuous	0,1,2,3,3.5
Dorm room number	Nominal	Discreet	0,1B,2,3,4,10,15,10A
Kelvin temperature	Ratio	Continuous	37,94,67,54.5

b)

Make	Mode
Fuel-type	Mode
# of doors	Mode, Median, binary discretization
Height	Mean, median, standard deviation, z-score normalization, Pearson's Correlation
# of Cylinders	Mode, Median
Price	Mean, median, standard deviation, z-score normalization, Pearson's Correlation

Assuming #of doors and # of cylinders to be ordinal attributes as the values can be compared.

c) We need to analyze the score of students. There can be two possible scenarios:
- All the scores would be discreet between 0-5. If this is the scenario, then we would consider the grades as an ordinal attribute. If all the grades are discreet, it would mean that the students never secure partial credits for a question and all the students will be distributed with discreet value between 0-5.
- The score of the students are continuous. If this is the scenario, then we would consider the grades to be a Ratio attribute. If the scores are continuous, it would mean that the students can secure partial credits for the question as will and the possibilities of the scores will be infinite between 0-5.

2.

a)

ID	Patient	Treatment	SBP
1	Patient 1	A	160
2	Patient 2	A	120
3	Patient 3	A	130
4	Patient 4	A	NA
5	Patient 5	A	120
6	Patient 6	A	NA
7	Patient 7	A	240
8	Patient 8	A	140
9	Patient 1	B	300
10	Patient 2	B	100
11	Patient 3	B	NA
12	Patient 4	B	130
13	Patient 5	B	110
14	Patient 6	B	100
15	Patient 7	B	120
16	Patient 8	B	90

b)

i) <u>Eliminating objects missing values:</u>
<u>Advantage:</u> Simple and effective strategy. This is the least time taking strategy to handle missing values.

<u>Disadvantage:</u> Loss of information. Even a partially specified data object contains some information and if many objects have missing values, then a reliable analysis can be difficult. Also, other attributes of those objects may have critical data.

ii)

Estimating the missing values:
Advantage: No loss of information incurred by eliminating data objects.

Disadvantage: This approach may lead to inconsistent bias and adds no new information but only increases the sample size and leads to an underestimate of the errors.

Based on the data above, I would choose estimation the missing values since the data set size is small and eliminating records with missing values will lead to loss of useful information.

c)

i) The results 240 and 300 are outliers because these values are unusual with respect to the typical values of that attribute.

Some of the strategies to handle outliers are:
Dropping the outlier records
Sometimes the outlier records can be removed from the dataset to prevent those records from skewing the analysis.

Assigning a new value
If an outlier seems to be due to mistakes during data collection, then replacing that value with the mean of the variable or using a regression model to predict the missing value could be tried.

Transforming the data
This approach involves creating a transformation of the data and using that instead of the original dataset.

ii) This inaccuracy creates noise as noise is the random component of a measurement error.
Some of the strategies to handle noise are:

Robust algorithms
Develop algorithms that produce acceptable results even when noise is present. This depends on the classification algorithm and therefore, the same result is not directly extensible to other learning algorithms, since the benefit comes from the adaptation itself. Moreover, this approach requires to change an existing method, which neither is always possible nor easy to develop.

Preprocessing
The datasets can be preprocessed to remove or correct the noisy records. However, this requires the usage of a preprocessing step, which is usually time

consuming. Furthermore, these methods are only designed to detect a specific type of noise and hence, the resulting data might not be perfect.

3.

a)

i) To determine the average salary of professors at NC State University, the faculty were divided into the following groups: instructors, assistant professors, associate professors, and professors. Twenty faculty members from each group were selected for the study.
Stratified sampling:
Stratified random sampling is a method of sampling that involves the division of a population into smaller groups known as strata and these are formed based on members' shared attributes or characteristics.
So, in the given problem, we have our population as professors at NC State University and the population is divided into smaller groups as instructors, assistant professors, associate professors, and professors based on some shared characteristics and the numbers of candidates selected from each group is not proportion to the group size but fixed to a size of 20. So, it is disproportionate stratification.

ii) From the following population, {2, 2, 4, 4, 6, 6, 8}, a sample {2,2,2,6,8} was collected.
Simple random sample with replacement:
2 is seen twice in the population, but it appears 3 times in the sample. This is more than what is available in the population from which sampling is done.
So, it should be Simple random sample with replacement.

iii) Data is collected in an experiment until a predictive model reaches 90% accuracy.
Adaptive sampling:
Adaptive sampling is where we adapt our selection criteria as the experiment progresses.
The sampling design is modified in real time as data collection continues – based on what has been learned from previous sampling.
In the given example also, we need to continue the process of data collection until we reach 90% with the model's prediction accuracy. So, we need to adapt our selection criteria as we progress to achieve the desired goal of reaching at least 90% prediction accuracy.

b)

i) Stratified random sampling is a method of sampling that involves the division of a population into smaller groups known as strata and these are formed based on members' shared attributes or characteristics.
In this given problem also, US people represent population, states represent strata and each chamber of US congress represents a sample here (as US has multiple states and members of every state share some common characteristics).

So, it makes sense to represent each state as Strata as it is important to have representatives from every state while doing survey or forming chambers of US congress. This stratification can be either proportionate or disproportionate.

ii) Senate Survey - As the chamber of Senates represents a sample created through disproportionate stratification (2 people from every state for a total of 100), It makes sense to send the survey to 20 members in Alaska (20 people from every state for a total of 1000).

iii) House Survey - As the chamber of House representatives represent a sample created through proportionate stratification, it makes sense to apply the same strategy for this survey also. So, the survey can be sent to some random 62 members of Florida (62 people from Florida for a total of 1000).

iv) **Senate Approach**(Disproportionate Stratification)**:**
US population count is very high in total numbers, but the Senate count is only 100. So, if we go with the approach of proportionate stratification, then the demands/characteristics of those states with high population would dominate over less populated states as the size of this sample (senate) is very small compared to the overall size of the population.
As a result, the views/demands of less populated states might go unnoticed and the characteristics of the overrepresented group can skew the overall results when we speak technically in terms of data analysis.
So, senate approach (disproportionate stratification) is an ideal one when the sample is too small compared to the overall population.

House Approach (Proportionate Stratification): This accurately reflects the population being studied (US population here) and it ensures that each subgroup (a state in US) receives proper and proportionate representation within the sample.
The ratio or distribution of the members from every state inside the sample (House) is proportional to the ratio of state's population to the overall population, and this type of sampling would exactly mimic the characteristics of the overall population.

4.

a) In PCA1, we should retain only principle component 1 as PC1 approximately captures all the variance in the raw data set. The contribution of other principle components to the variance is approximately 0.

b) In PCA1, $petal.width$ explains the most variance. The features $petal.width$ and $petal.length$ contribute to the majority of the variance in the dataset.

c) In PCA2, we should retain principle component 1 and principle component 2 as PC1 and PC2 approximately captures all the variance in the raw data set. The contribution of other principle contribution to the variance is approximately 0.

d) In PCA2, according to first principle component, all the features have significant contribution in the variance of the data set. Features namely $sepal.length$, $sepal.width$, $petal.length$, $petal.width$.

e) PCA1 is done on the data set without normalizing it and PCA2 is performed after normalizing the data set. PCA projects the data onto the direction which will maximize the variance. As PCA maximizes the variance, normalization is an important step before performing the PCA. If the variance for some features in the data set are very high compared to others, these features will dominate in the PCA if the normalization is not performed. To get an unbiased result, we should normalize the data before performing PCA. Hence, we should choose PCA2 for our analysis.

f) Based on the results of PCA1 and PCA2 we can conclude that all the features in the data set have significant contribution towards the variance. Therefore, we should include all the features for further analysis.

5.

a) Range of Temperature attribute = 95-50 = 45
Interval width = 45/4 = 11.25

Intervals:
[50.0 - 61.25) = I1
[61.25 – 72.5) = I2
[72.5 – 83.75) = I3
[83.75 – 95.0] = I4

No.	OUTLOOK	TEMPERATURE	HUMIDITY	WINDY	PLAY
1	sunny	I4	85.0	FALSE	no
2	sunny	I3	90.0	TRUE	no
3	overcast	I3	86.0	FALSE	yes
4	rainy	I2	96.0	FALSE	yes
5	rainy	I2	80.0	FALSE	yes
6	rainy	I2	70.0	TRUE	no
7	overcast	I2	65.0	TRUE	yes
8	sunny	I2	95.0	FALSE	no
9	sunny	I2	70.0	FALSE	yes
10	rainy	I3	80.0	FALSE	yes
11	sunny	I3	71.0	TRUE	yes
12	overcast	I3	89.0	TRUE	yes
13	overcast	I3	75.0	FALSE	yes
14	rainy	I2	91.0	TRUE	no
15	sunny	I4	85.0	FALSE	yes
16	rainy	I1	45.0	YES	no

b) Binning by ensuring that each interval gets equal number of entries.
Intervals:

[45.0 - 70.00] = I1
[71.0 – 80.0] = I2
[81.0 – 89.00] = I3
[90.00 – 96.0] = I4

No.	OUTLOOK	TEMPERATURE	HUMIDITY	WINDY	PLAY
1	sunny	85.0	I3	FALSE	no
2	sunny	80.0	I4	TRUE	no
3	overcast	83.0	I3	FALSE	yes
4	rainy	70.0	I4	FALSE	yes
5	rainy	68.0	I2	FALSE	yes
6	rainy	65.0	I1	TRUE	no
7	overcast	64.0	I1	TRUE	yes
8	sunny	72.0	I4	FALSE	no
9	sunny	69.0	I1	FALSE	yes
10	rainy	75.0	I2	FALSE	yes
11	sunny	75.0	I2	TRUE	yes
12	overcast	73.0	I3	TRUE	yes
13	overcast	81.0	I2	FALSE	yes
14	rainy	71.0	I4	TRUE	no
15	sunny	95.0	I3	FALSE	yes
16	rainy	50.0	I1	YES	No

c) Intervals:
Finding out all the required intervals by using different values for 'k' in the given formula to represent the humidity data given in the table.

'k' values used = -3, -2, -1, 0, 1, 2

[41.0 – 54.00) = I1
[54.0 – 67.00) = I2
[67.0 – 80.0) = I3
[80.0 - 93.00) = I4
[93.00 – 106.0) = I5

No.	OUTLOOK	TEMPERATURE	HUMIDITY	WINDY	PLAY
1	sunny	85.0	I4	FALSE	no
2	sunny	80.0	I4	TRUE	no
3	overcast	83.0	I4	FALSE	yes
4	rainy	70.0	I5	FALSE	yes
5	rainy	68.0	I4	FALSE	yes
6	rainy	65.0	I3	TRUE	no
7	overcast	64.0	I2	TRUE	yes
8	sunny	72.0	I5	FALSE	no
9	sunny	69.0	I3	FALSE	yes
10	rainy	75.0	I4	FALSE	yes
11	sunny	75.0	I3	TRUE	yes
12	overcast	73.0	I4	TRUE	yes
13	overcast	81.0	I3	FALSE	yes
14	rainy	71.0	I4	TRUE	no
15	sunny	95.0	I4	FALSE	Yes
16	rainy	50.0	I1	YES	No

d) **Equal-width:**

Advantages:

Most straight forward approach – simple and easy to implement.

Produces a reasonable abstraction of data, good for many classes.

Disadvantages:

Can fail miserably for unequal distributions.

How many bins to use? – not so easy to decide.

Different bins will have different number of data points and some bins might remain empty only.

Very sensitive to outliers.

Skewed data is not handled well.

When to use:

When the data is equally distributed and not skewed.

When the dataset is free from outliers or has minimal outliers.

When we want to go with a simple way of binning.

Equal-depth:

Advantages:

Gives better results compared to equal-width approach.

Insensitive to outliers.

Bins empty situation will never occur.

Results in good data scaling.

Disadvantages:

How many bins to use? – not so easy to decide.

Managing categorical features can be tricky.

When to use:

When we have outliers in the dataset.

When the dataset is skew d or has unequal distributions.

Third approach

This is similar to equal-width approach and the interval size here is decided based on the standard deviation value.

This approach has same advantages and disadvantages as equal-width approach but might act slightly better than the equal-width approach as it includes standard deviation and mean (statistic measures) terms while calculating the size (interval) of the bins.

When to use:

When the data is equally distributed and not skewed.
When the dataset is free from outliers or has minimal outliers.
When we want to go with a simple way of binning.

6.

a)

i) Euclidean distance

Euclidean distance is defined by

$$d(x,y) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2}$$

Positive Definiteness:

Euclidean distance satisfies this property because, as it can be seen from the above equation, this distance calculation involves a summation of squares.

$$d(x,y) \geq 0 \; for \; any \; x \; and \; y$$
$$and, d(x,y) = 0 \; if \; and \; only \; if \; x = y$$

Symmetry:

$$d(x,y) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2} = \sqrt{\sum_{k=1}^{n}(y_k - x_k)^2} = d(y,x)$$

$Therefore$, Euclidean distance satisfies the symmetry property.

Triangle Inequality:

It is a mathematical property of triangles that the length of the third side is always less than or equal to the sum of the length of the other two sides. Consider three vectors x, y, z to be the vertices of a triangle. In this case, Euclidean distance actually calculates the length of each side.

$$Therefore, d(x,z) \leq d(x,y) + d(y,z)$$

Hence Euclidean distance satisfies the triangle inequality.

ii) Manhattan distance:

Manhattan distance is defined as

$$d(x,y) = \sqrt{\sum_{k=1}^{n} |x_k - y_k|}$$

Positive definiteness:

As Manhattan distance calculation involves a summation of absolute values, the distance is always going to be positive with the exception that the distance will equal zero if and only if $x = y$. Therefore, Manhattan distance satisfies the positive definiteness property.

$d(x,y) \geq 0 \; for \; any \; x, y$

$d(x,y) = 0 \; iff \; x = y$

Symmetry:

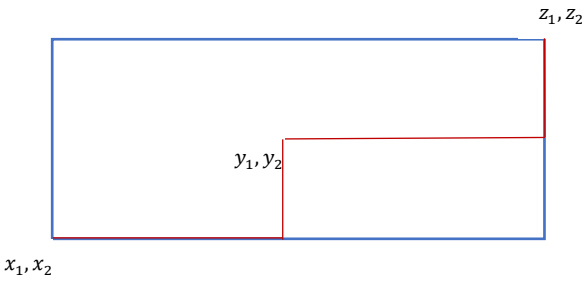
$$d(x,y) = \sqrt{\sum_{k=1}^{n} |x_k - y_k|} = \sqrt{\sum_{k=1}^{n} |y_k - x_k|} = d(y,x)$$

$Therefore$, Manhattan distance is symmetric.

Triangle Inequality:

Manhattan distance satisfies triangle inequality. For justification, consider three vectors x, y, z , each having two dimensions.

In this case, as can be seen from the fig shown below, for any y, inside and on the box, $d(x,z) = d(x,y) + d(x,z)$ and for all other y, $d(x,z) < d(x,y) + d(y,z)$.



$$Therefore, d(x,z) \leq d(x,y) + d(y,z)$$

iii) Divergent function defined between two sets

$$d(A,B) = 1 - \frac{|A \cap B|}{|A|}$$

Positive definiteness:

$d(A,B) = 0$ doesn't require both the sets to be the same. For example, let

$$A = \{1,2\}\ and\ B = \{1,2,3,4\}$$

$$d(A,B) = 1 - \frac{2}{2} = 0$$

Hence, this distance function doesn't satisfy the positiveness property.

Symmetry:

This distance function is not symmetric because $d(A,B)$ represents the number of A's items not present in B where as $d(B,A)$ represents number of B's items not present in A.

For example:

Let, $A = \{1,2\}\ and\ B = \{1,2,3,4\}$

$$d(A,B) = 1 - \frac{|A \cap B|}{|A|} = 1 - \frac{2}{2} = 0$$

$$d(B,A) = 1 - \frac{|B \cap A|}{|B|} = 1 - \frac{2}{4} = 0.5$$

Triangle Inequality:

This distance function does not satisfy the triangle inequality. This could be shown by a counter example.

Let,

$$A = \{1,2\}$$

$$B = \{1,2,3,4,5,6,7,8\}$$

$$C = \{3,4\}$$

$$d(A,C) = 1 - \frac{|A \cap C|}{|A|} = 1 - \frac{0}{2} = 1$$

$$d(A,B) = 1 - \frac{|A \cap B|}{|A|} = 1 - \frac{2}{2} = 0$$

$$d(B,C) = 1 - \frac{|B \cap C|}{|B|} = 1 - \frac{2}{8} = 0.75$$

$$d(A,B) + d(B,C) = 0.75$$

$$Therefore, d(A,C) > d(A,B) + d(B,C)$$

iv) Cosine distance:

Cosine distance between two numeric vectors is defined as

$$d(A,B) = 1 - \frac{A.B}{||A||.||B||} = 1 - cos(\theta),$$

$where\ \theta\ is\ the\ angle\ between\ the\ vectors.$

Positive definiteness:

$d(x,y) = 0\ doesn't\ require\ x = y$, for $cosine\ distance$. For example,
Let, $x = [1,1,1] and\ y = [2,2,2]$ be two non-equal vectors.

$$d(x,y) = 1 - \frac{2\times1 + 2\times1 + 2\times1}{\sqrt{2^2+2^2+2^2}\times\sqrt{1^2+1^2+1^2}} = 0$$

Hence, $cosine\ distance$ doesn't satisfy the positive definiteness property.

Symmetry:

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

$$= d(B,A)$$

Hence the $cosine\ distance$ is symmetric.

Triangle Inequality:

Cosine distance doesn't satisfy triangle inequality. This could be shown by a counter example. Let

$A = [1,0], B = \left[\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right], C = [0,1]$ be three vectors.

$$d(A,C) = 1 - \frac{1\times0 + 0\times1}{\sqrt{1^2+0^2}\times\sqrt{0^2+1^2}} = 1$$

$$d(A,B) = 1 - \frac{1\times\frac{\sqrt{2}}{2} + 0\times\frac{\sqrt{2}}{2}}{\sqrt{1^2+0^2}\times\sqrt{\left(\frac{\sqrt{2}}{2}\right)^2+\left(\frac{\sqrt{2}}{2}\right)^2}} = 1 - \frac{\sqrt{2}}{2}$$

$$d(B,C) = 1 - \frac{\frac{\sqrt{2}}{2}\times0 + \frac{\sqrt{2}}{2}\times1}{\sqrt{\left(\frac{\sqrt{2}}{2}\right)^2+\left(\frac{\sqrt{2}}{2}\right)^2}\times\sqrt{1^2+0^2}} = 1 - \frac{\sqrt{2}}{2}$$

$$d(A,B) + d(B,C) = 1 - \frac{\sqrt{2}}{2} + 1 - \frac{\sqrt{2}}{2} = 2 - \sqrt{2} \approx 0.586$$

$$\therefore d(A,C) > d(A,B) + d(B,C)$$

b)

i) The triangle inequality as well as the positive definiteness property can be used to skip comparisons.

ii) Let $x1, x2 \in X \; and \; y \in Y$. Suppose $d(x1, y)$ was already calculated. By the triangle inequality,

$$d(x1, y) \leq d(x1, x2) + d(x2, y)$$
$$\rightarrow d(x2, y) \geq d(x1, y) - d(x1, x2)$$

Lower bound for $d(x2, y)$, $\delta = d(x1, y) - d(x1, x2)$
If the lower bound, δ is greater than $d(y, x1)$, then the calculating $d(y, x2)$ can be skipped.

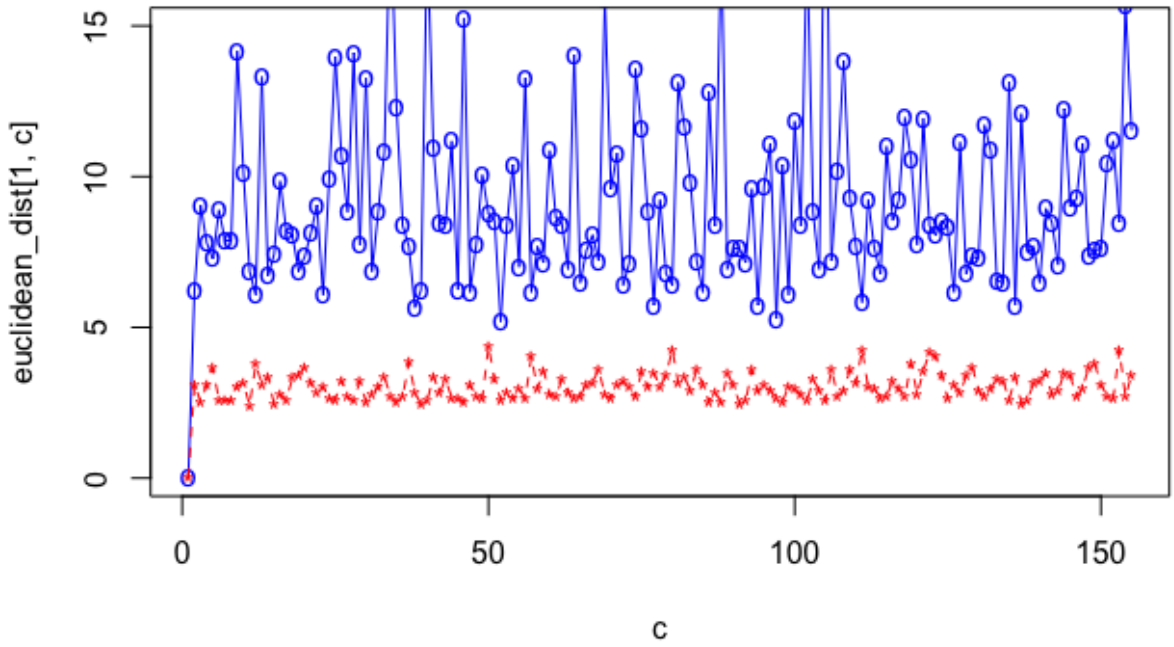
Also, we can also make use of the $d(x, y) \geq 0$ property of the distance metric to skip comparisons. If any distance calculation returns the distance as 0, then all the other calculations can be skipped because that is the lowest possible distance.

iii) This strategy does reduce the number of comparisons in the best case. The best case is when the minimum possible distance is calculated first. This will result in doing just one $d(x, y)$ calculation and skipping all the rest. However, in the worst case, when the points in X are arranged in the decreasing order of their distances from y, none of the $d(x, y)$ calculations are skipped.

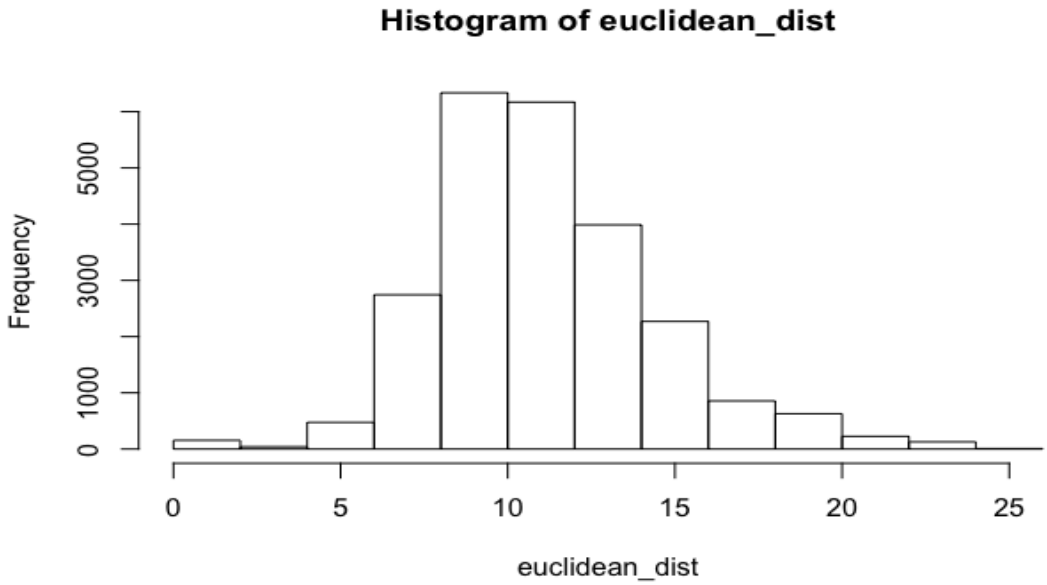
7. Changes identified between the original and the normalized distances.

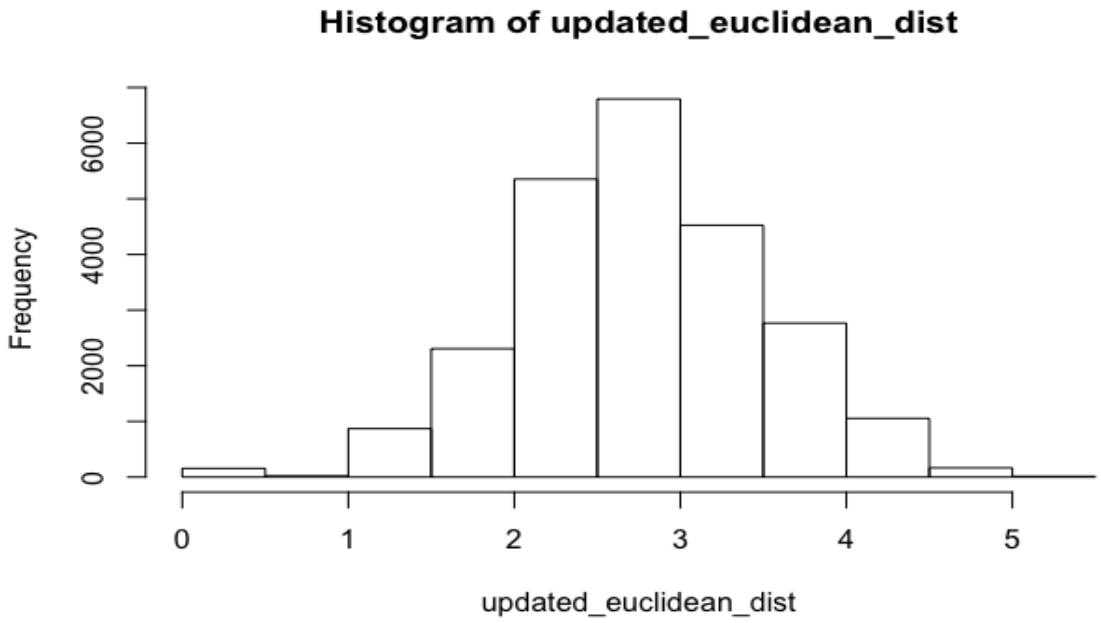
i) The dissimilarity between the vectors is much more pronounced in the original distance calculation than in the normalized distance calculation. This could be because of the fact that when different variables are combined, in their raw form, some variables could dominate over other variables and affect the results of the calculations.

ii) No outliers. This is because of the fact that before normalization, there were no boundaries for the data and normalizing produced boundaries.



iii) The histogram of the original distance matrix is skewed to the left whereas the histogram of the normalized distance appears more like a normal distribution.




Histogram of updated_euclidean_dist
Frequency
0
2000
4000
6000
0
1
2
3
4
5
updated_euclidean_dist