SIMILARITY MEASURE

The scripts are classified into five parts

1. 0_Decide the number of bins.R :
2. 1_Binning.R:
3. 2_Filtering.R:
4. 3_Euclidean_distance.R
5. main.R

Unzip the package and place "observation.txt" file in this folder. "observation.txt" file is a semi-colon separated file containing data of all the patients for all attributes.

In order to execute the main script, input the following commands in the R shell:

```
source("<<Path to the extracted R_package>>/code/main.R")
```

```
similarity_measure(<<Path to the extracted R_package>>,
<<Name of the observation file>>, <<index_patient_id>>,
<<index_attribute>>)
```

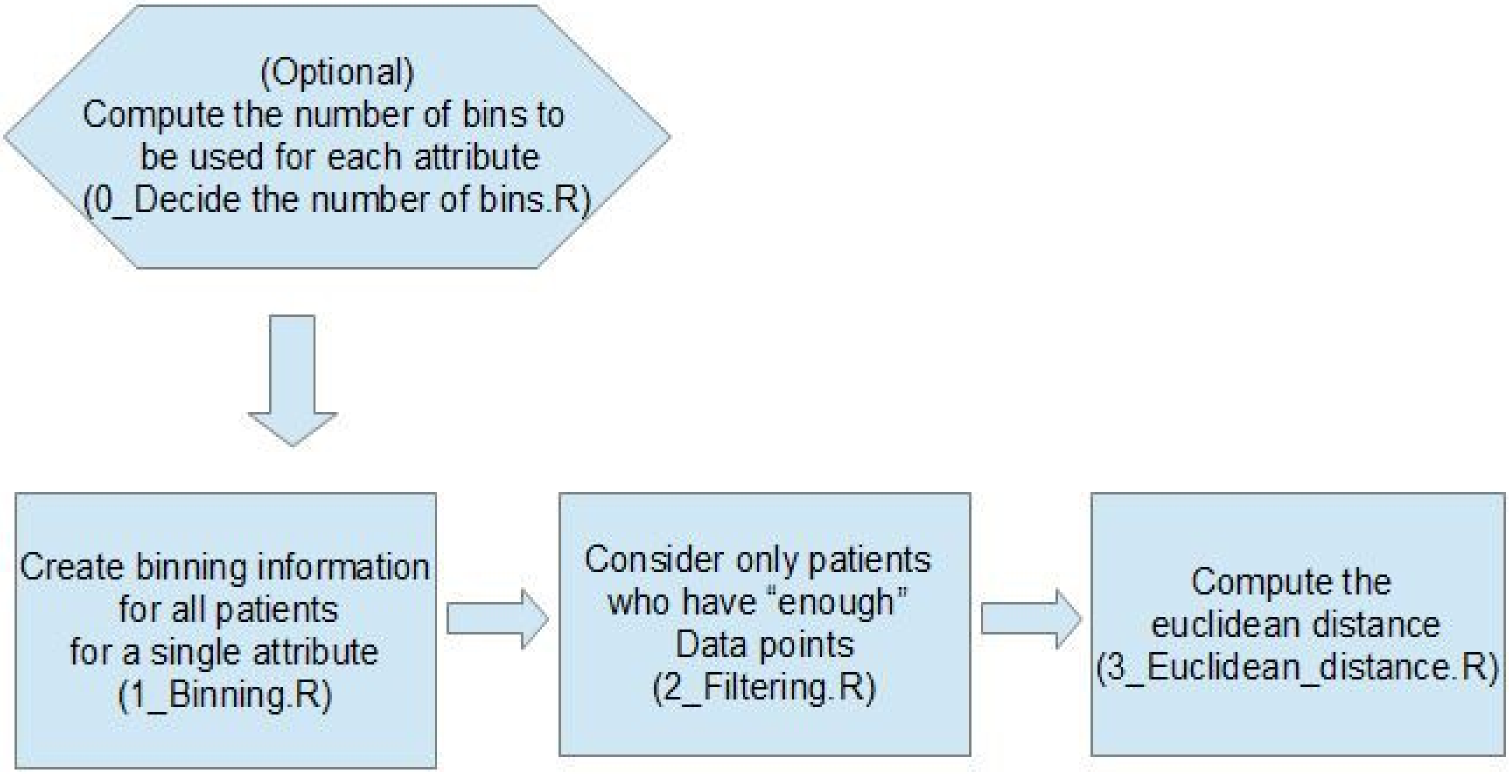
Sample valid argument values:

```
index_patient_id=138312
index_attribute=15  #Heart Rate (Refer concept.txt)
```

Example:

```
source("E:/UTAH/test/code/main.R")
```

```
similarity_measure("E:/UTAH/test", "observation.txt", 138312,
15)
```


(Optional)
Compute the number of bins to be used for each attribute
(0_Decide the number of bins.R)
Create binning information for all patients for a single attribute
(1_Binning.R)
Consider only patients who have “enough” Data points
(2_Filtering.R)
Compute the euclidean distance
(3_Euclidean_distance.R)

1. 0_Decide the number of bins.R :

This is a one time execution which decides the number of bins to be used for each of the 37 attributes based on the number of observations recorded for that attribute across all patients.

Input	Output
observation.txt	Binning_Data/Number_of_bins.csv

The output here is just a suggestion on the bin size to be used based on the data.

One can manually enter the "number of bins" required for each attribute in the file "Number_of_bins.csv".

One can also use "Suggested_Number_of_bins.csv" where we have given some thought to each attribute and suggested the number of bins to be used based on our judgement. Make sure to rename this file to "Number_of_bins.csv"

2. 1_Binning.R:

This is a one time execution for a particular attribute. This creates the bins and stores the bin values.

Input	Output
observation.txt	Bins.csv files containing the binning information
Number_of_bins.csv	
index_attribute	

For example, index_attribute=15 (Heart rate), this would create a directory named "15_bins". This folder would contain 4000 csv files with the patient ID as its name (like 138312.csv)

Once this is created, any query with a patient ID can be initiated for this particular attribute. In case the "Number_of_bins.csv" is modified, then this script should be executed again

3. 2_Filtering.R:

Remove patients who have very few observations for a particular attribute

Input	Output
index_patient_id	Filtered list of patients
index_attribute	
bins created by 1_Binning.R	

Lets say N = Number of non empty bins of index patient,
Then, we will consider only patients who have at least N/2 non empty bins

4. 3_Euclidean_distance.R

Compute the Normalized Euclidean Distance between the index patient and every patient in the "filtered list of patients".

Input	Output
index_patient_id	Euclidean_Distance.csv
index_attribute	
Bins created by 1_Binning.R	
filtered list of patients	

Example:

Let, Number of bins = 16
Number of overlapping (intersection) non-empty bins of Patient A and Patient B = 14

Then we compute the square of the Euclidean distance for the overlapping bins and divide by 14.