NAVEEN BALARAJU

Goal: To perform SOM analysis on Wisconsin Breast cancer dataset.

Dataset description: This breast cancer databases is obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg and it consists of 699 observations and 11 variables

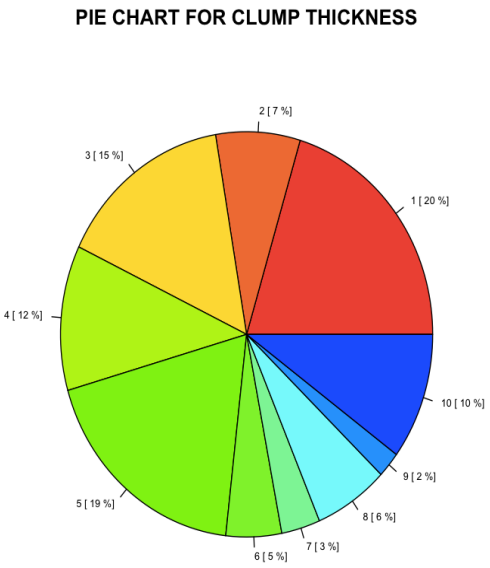
1. Sample code number: id number
2. Clump Thickness: 1 - 10
3. Uniformity of Cell Size: 1 - 10
4. Uniformity of Cell Shape: 1 - 10
5. Marginal Adhesion: 1 - 10
6. Single Epithelial Cell Size: 1 - 10
7. Bare Nuclei: 1 - 10
8. Bland Chromatin: 1 - 10
9. Normal Nucleoli: 1 - 10
10. Mitoses: 1 - 10
11. Class_type (2 for benign, 4 for malignant)

Exploratory Data Analysis:

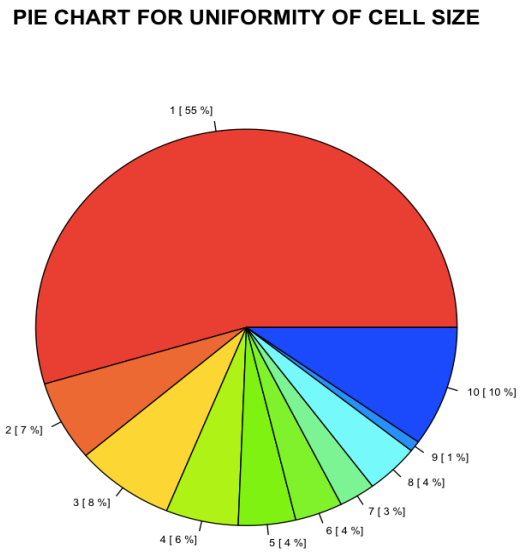
Data Preprocessing: Since the raw dataset did not contain column names, I've renamed the columns according to the Data description given in the sources. Furthermore, "Bare Nuclei" contains 16 observations with unknown levels indicated by "?", I've reordered the levels from 1 through 10 and deleted those missing observations.

Data Visualization using Pie charts:

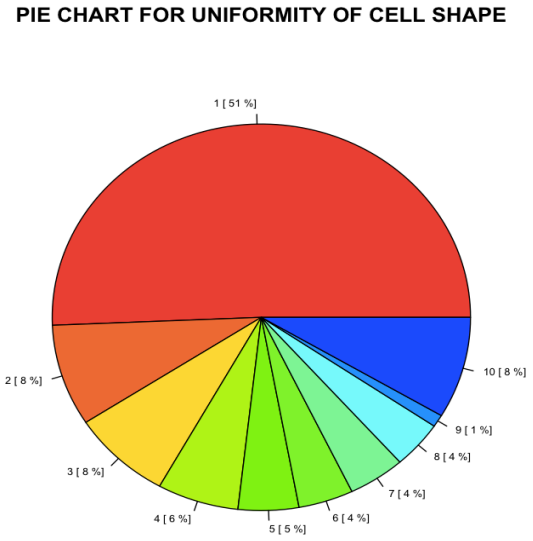
1. <u>Analysis of "clump thickness":</u> From the pie chart I can infer that 20% of the observations have clump thickness of 1,7% have thickness of 2,15% have thickness of 3,12% have thickness of 4,19% have thickness 5,5% have thickness of 6,3% have thickness 7,6% have thickness 8,2% have thickness 9 and the rest 10% have clump thickness of 10.



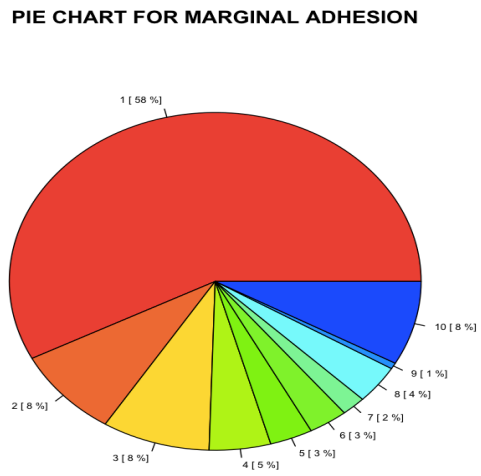
2. Analysis of “uniformity of cell size”: From the Pie chart I can infer that 55% of the observations have uniformity cell size of 1,7% have cell size of 2,8% have cell size 3,6% have cell size 4, 4% have cell size cell size 5, 4% have cell size 6,3% have cell size 7,4% have cell size 8,1% have cell size 9 and rest 10% have size of 10.



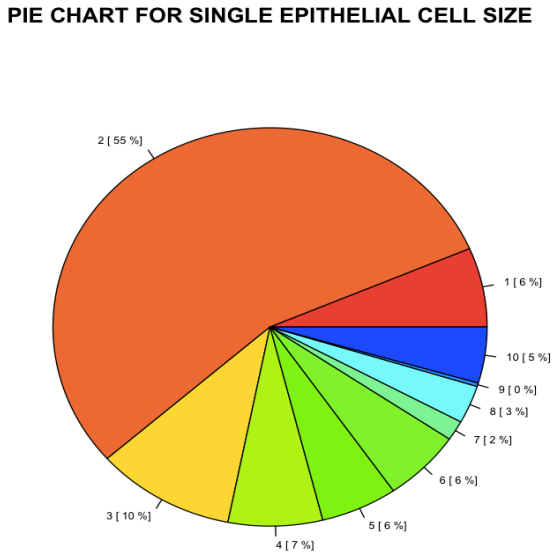
3. Analysis of “uniformity of cell shape”: From the Pie chart I can infer that 51% of the observations have uniformity cell Shape of 1,8% have cell shape of 2,8% have cell shape of 3,6% have cell shape of 4, 5% have cell shape of 5, 4% have cell shape of 6,4% have cell shape of 7,4% have cell shape of 8,1% have cell shape of 9 and rest 8% have uniformity cell shape of 10.



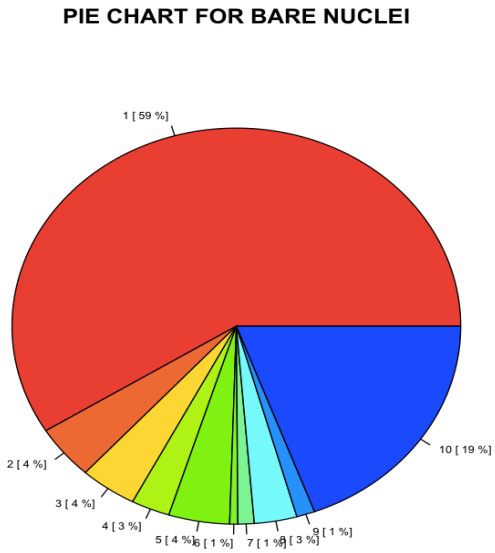
4. Analysis of “Marginal Adhesion”: From the Pie chart I can infer that 58% of the observations have Marginal Adhesion (MA) of 1,8% have MA of 2,8% have MA of 3,5% have MA of 4, 3% have MA of 5, 3% have MA of 6, 2% have MA of 7,4% have MA of 8,1% have MA of 9 and rest 8% have Marginal Adhesion (MA) of 10.



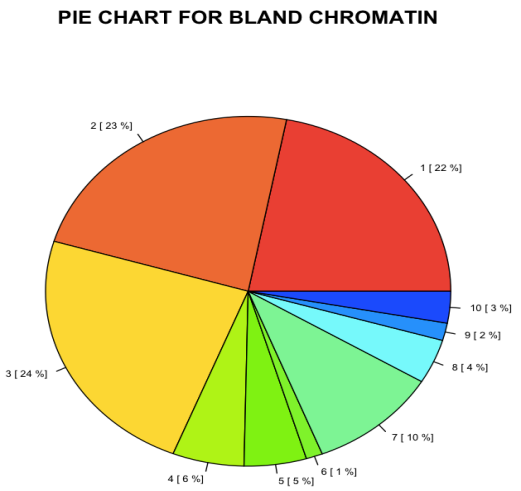
5. Analysis of “Single Epithelial cell size”: From the Pie chart I can infer that 6% of the observations have Single Epithelial cell size of 1,55% have cell size of 2,10% have cell size of 3,7% have cell size of 4, 6% have cell size of 5, 6% have cell size of 6,2% have cell size of 7,3% have cell size of 8 and 5% have Single Epithelial cell size of 10.



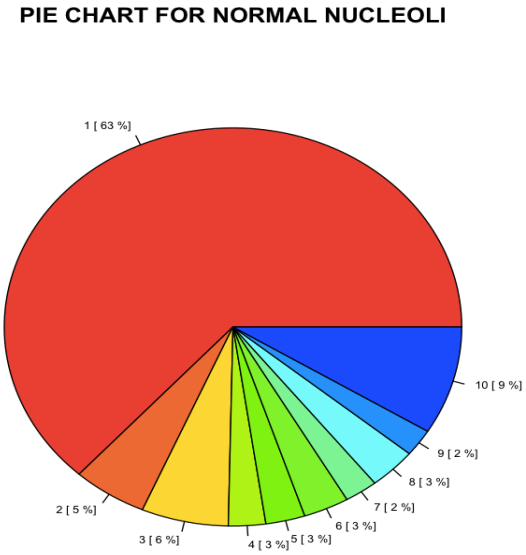
6. Analysis of “Bare Nuclei”: From the Pie chart, I can infer that 59% of the observation have Bare Nuclei (BN) size of 1,4% have BN of 2,4% have BN of 3, 3% have BN of 4,4% have BN of 5, 1% have BN of 6,1% have BN of 7,3% have BN of 8,1% have BN of 9 and rest 19% have Bare Nuclei (BN) of 10.



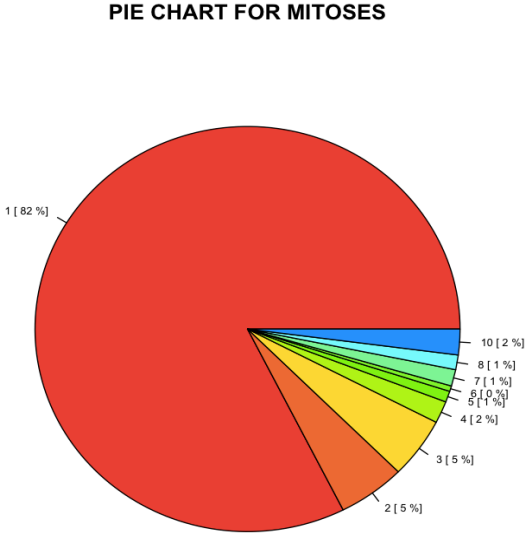
7. Analysis of “Bland Chromatin”: From the Pie chart, I can infer that 22% of the observation have Bland Chromatin (BC) of 1,23% have BC of 2,24% have BC 3, 6% have BC of 4,5% have BC of 5, 1% have BC of 6,10% have BC of 7,4% have BC of 8, 2% have BC of 9 and rest 3% have Bland Chromatin (BC) of 10.



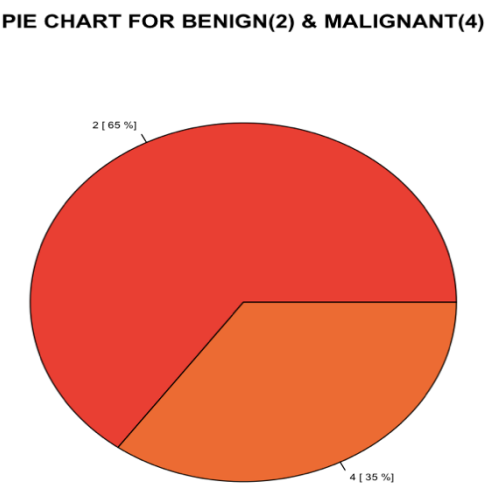
8. Analysis of “Normal Nucleoli”: From Pie chart, I can infer that 63% of the observations have Normal Nucleoli (NN) size of 1, 5% have NN of 2,6% have NN 3,3% have NN of 4,3% have NN of 5, 3% have NN of 6, 2% have NN of 7, 3% have NN of 8,2% have NN of 9 and the rest 9% have Normal Nucleoli (NN) of size 10.



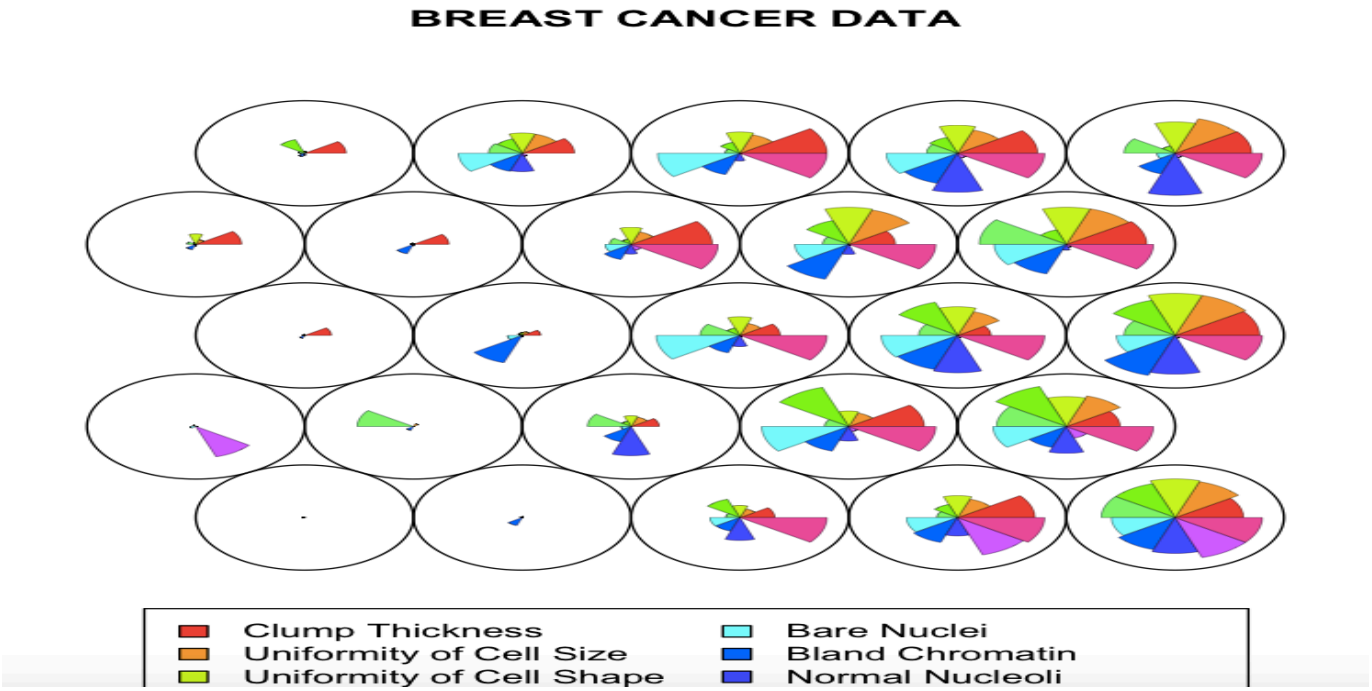
9. Analysis of “Mitoses”: From the Pie chart, I can infer that 82% of the observations have the Mitoses value of 1, while the rest 19% of the observations have other values for mitoses.



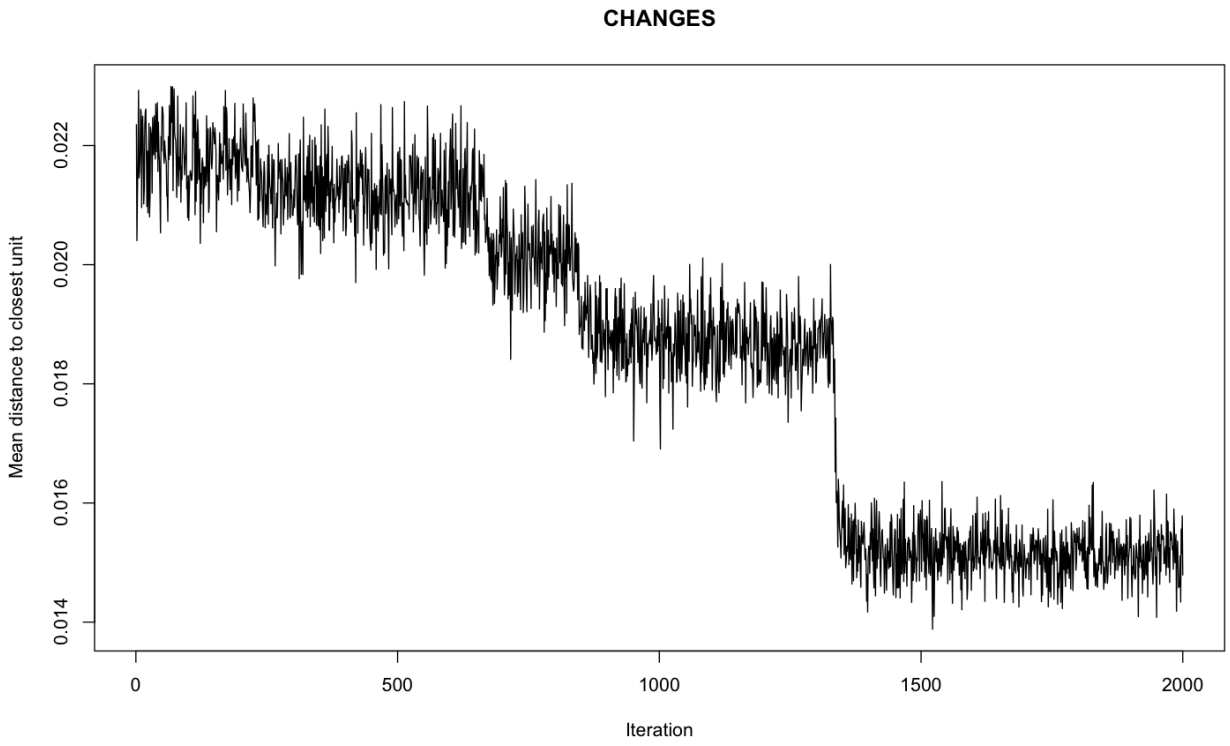
10. Analysis of “class type”: From the pie chart, I can infer that 65% of the observation are Benign and rest 35% is malignant.



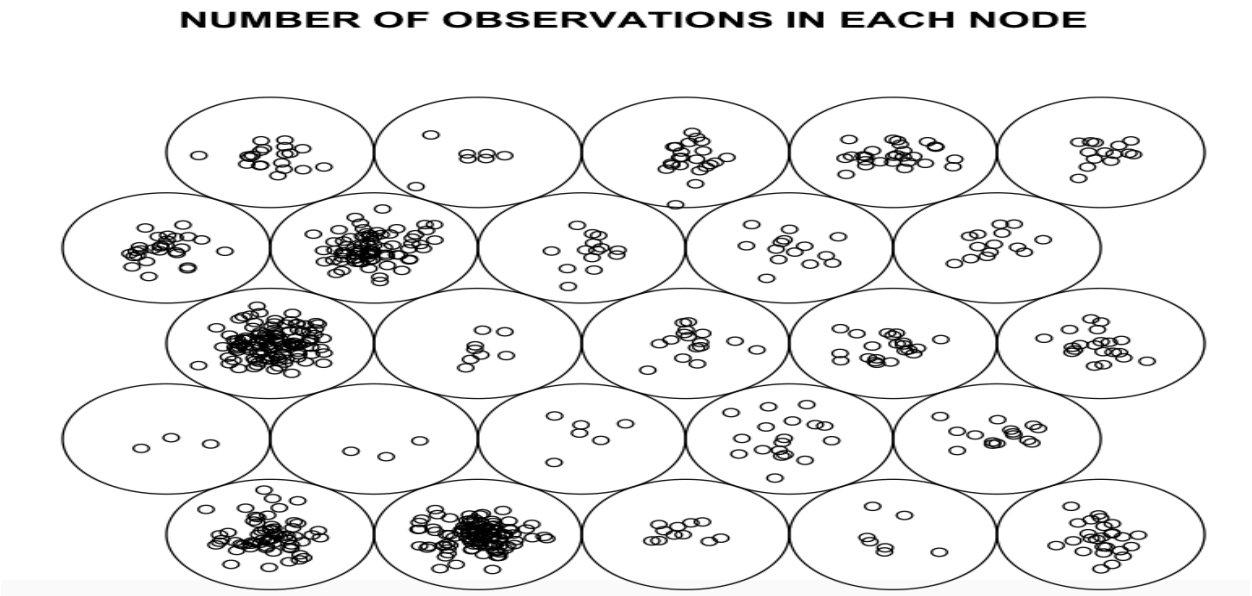
Plot of Code Book factors:



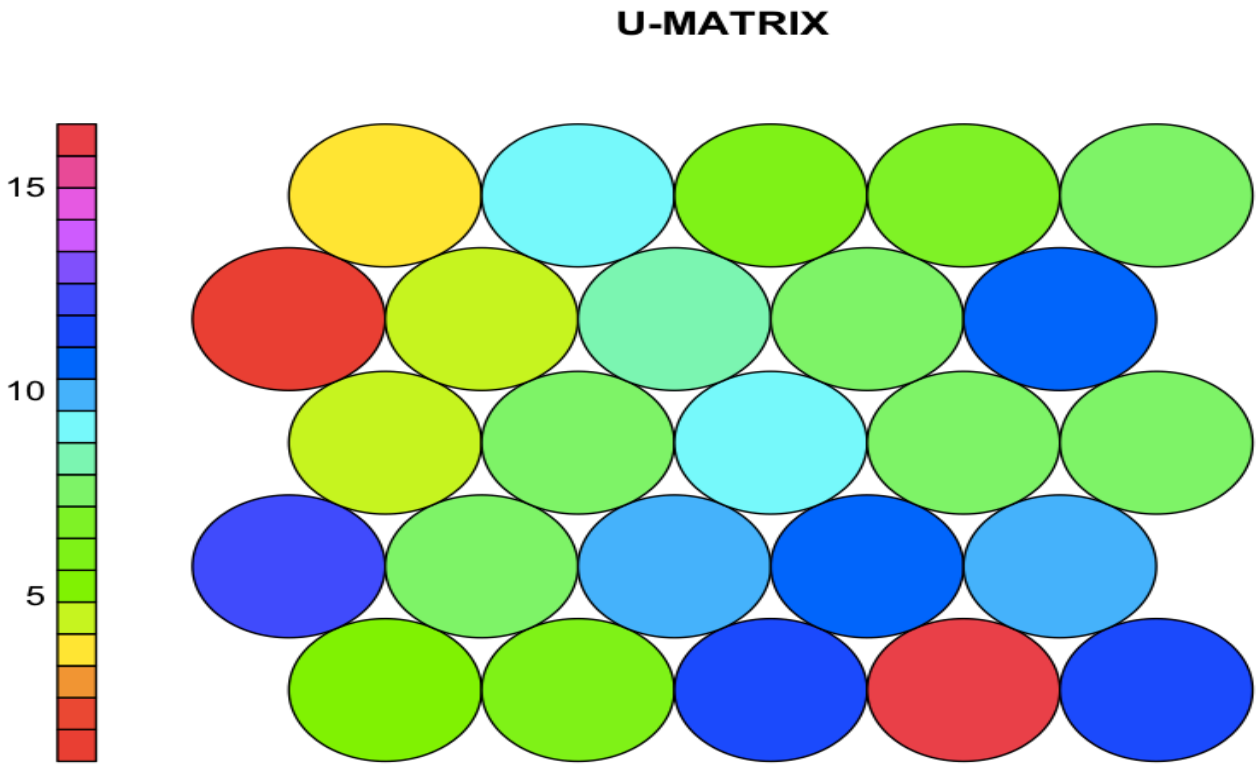
Plot of Mean distance to closest unit:



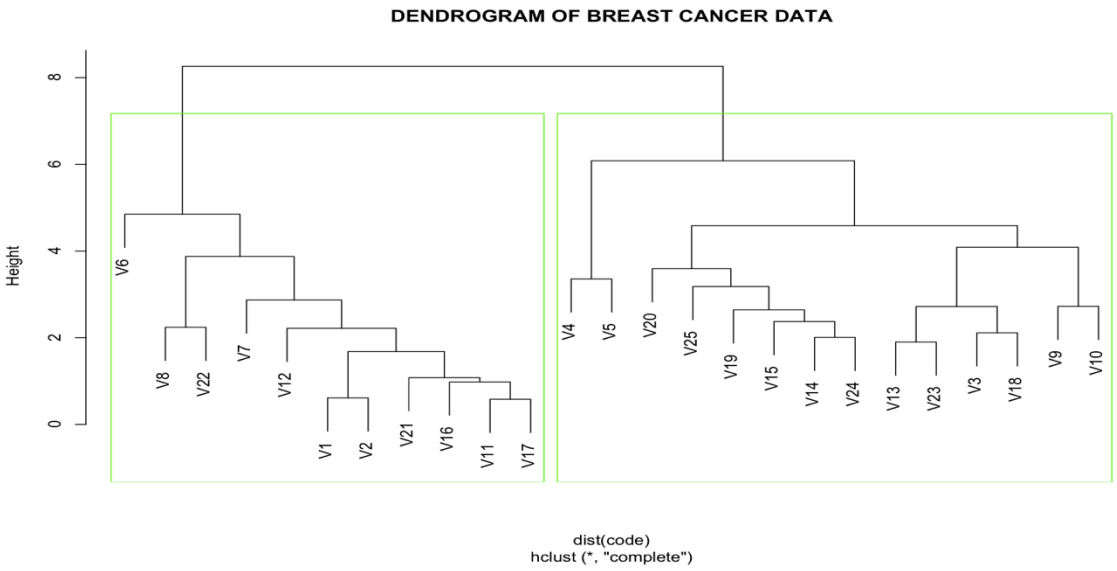
Plot of number of observations in each node:



Plot of Umatrix:



Hierarchical clustering: performing hierarchical clustering on the codes generated from Batch SOM analysis and cutting the dendrogram at h=6.5 results in two healthy looking clusters representing the Benign and Malignant cancer.



Plotting SOM with clusters found from hierarchical clustering: plotting SOM using the results obtained from hierarchical clusters produces a good demarcation between the Malignant (BLUE) and Benign (GREEN)