NORTHWESTERN UNIVERSITY | SCHOOL OF PROFESSIONAL STUDIES

Assignment #3: Multidimensional Scaling and Self-Organizing Maps

MSDS 411

This assignment has two components parts, one for each of these methods. They use different data sets. Please label your write-up appropriately to keep these analyses separate. The components are worth equivalent number of points.

COMPONENT 1: Multidimensional Scaling

In this component of the assignment, you will be conducting MDS analyses.

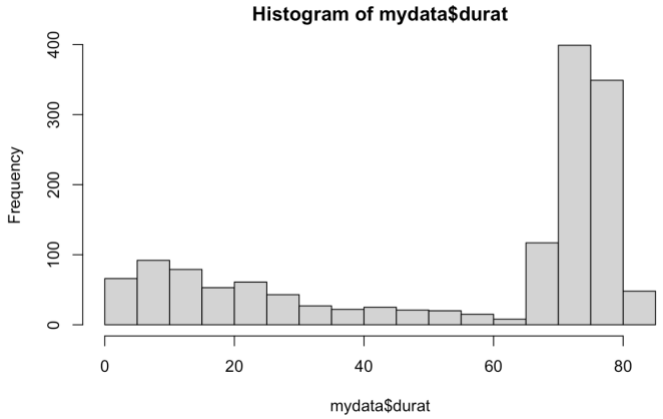
Data: The RECIDIVISM dataset is an 18-variable dataset with n=1445 records. Please see the data description file for the variable definitions and additional information about the dataset. The data consists of a random sample records on convicts released from prison during 1977/1978.

Assignment Tasks

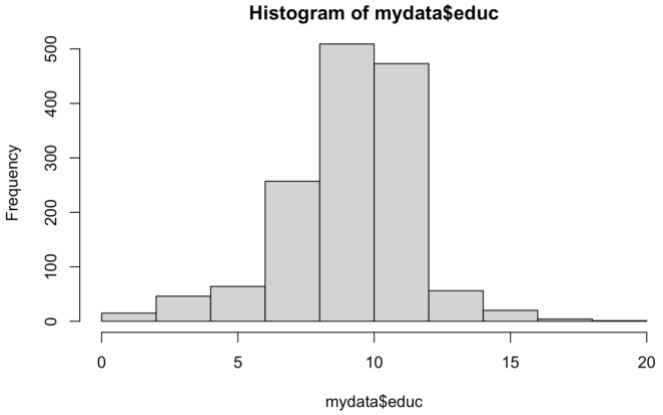
1) Perform a basic Exploratory Data Analysis on the Recidivism data. Report what you have learned through this activity. Prepare the data as best you can for an upcoming MDS analysis.

```r
summary(mydata)
```

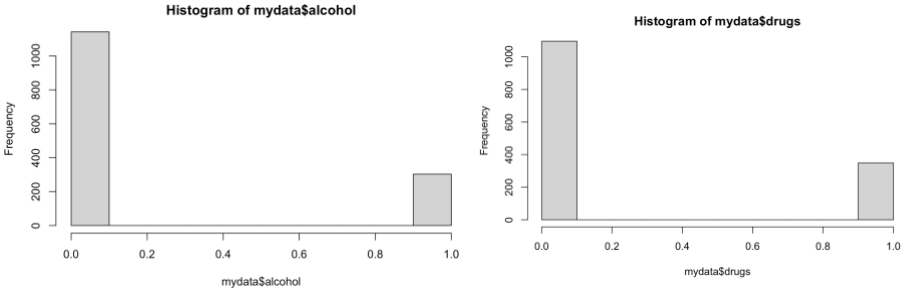
```
     black            alcohol           drugs             super            married           felon             workprg      
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000  
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000  
 Median :0.0000   Median :0.0000   Median :0.0000   Median :1.0000   Median :0.0000   Median :0.0000   Median :0.0000  
 Mean   :0.4851   Mean   :0.2097   Mean   :0.2415   Mean   :0.6941   Mean   :0.2554   Mean   :0.3142   Mean   :0.4651  
 3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000  
 Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000  
    property          person            priors            educ            rules             age             tserved      
 Min.   :0.0000   Min.   :0.00000   Min.   : 0.000   Min.   : 1.000   Min.   : 0.000   Min.   :198.0   Min.   :  0.00  
 1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.: 0.000   1st Qu.: 8.000   1st Qu.: 0.000   1st Qu.:258.0   1st Qu.:  6.00  
 Median :0.0000   Median :0.00000   Median : 0.000   Median :10.000   Median : 0.000   Median :307.0   Median : 12.00  
 Mean   :0.2547   Mean   :0.05329   Mean   : 1.432   Mean   : 9.702   Mean   : 1.185   Mean   :345.4   Mean   : 19.18  
 3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.: 2.000   3rd Qu.:11.000   3rd Qu.: 1.000   3rd Qu.:395.0   3rd Qu.: 25.00  
 Max.   :1.0000   Max.   :1.00000   Max.   :28.000   Max.   :19.000   Max.   :27.000   Max.   :933.0   Max.   :219.00  
     follow           durat             cens            ldurat      
 Min.   :70.00   Min.   : 1.00   Min.   :0.000   Min.   :0.000  
 1st Qu.:72.00   1st Qu.:27.00   1st Qu.:0.000   1st Qu.:3.296  
 Median :74.00   Median :71.00   Median :1.000   Median :4.263  
 Mean   :74.89   Mean   :55.37   Mean   :0.618   Mean   :3.745  
 3rd Qu.:78.00   3rd Qu.:76.00   3rd Qu.:1.000   3rd Qu.:4.331  
 Max.   :81.00   Max.   :81.00   Max.   :1.000   Max.   :4.394  
```



Lots of these criminals are in the high duration levels. This distribution is left skewed.



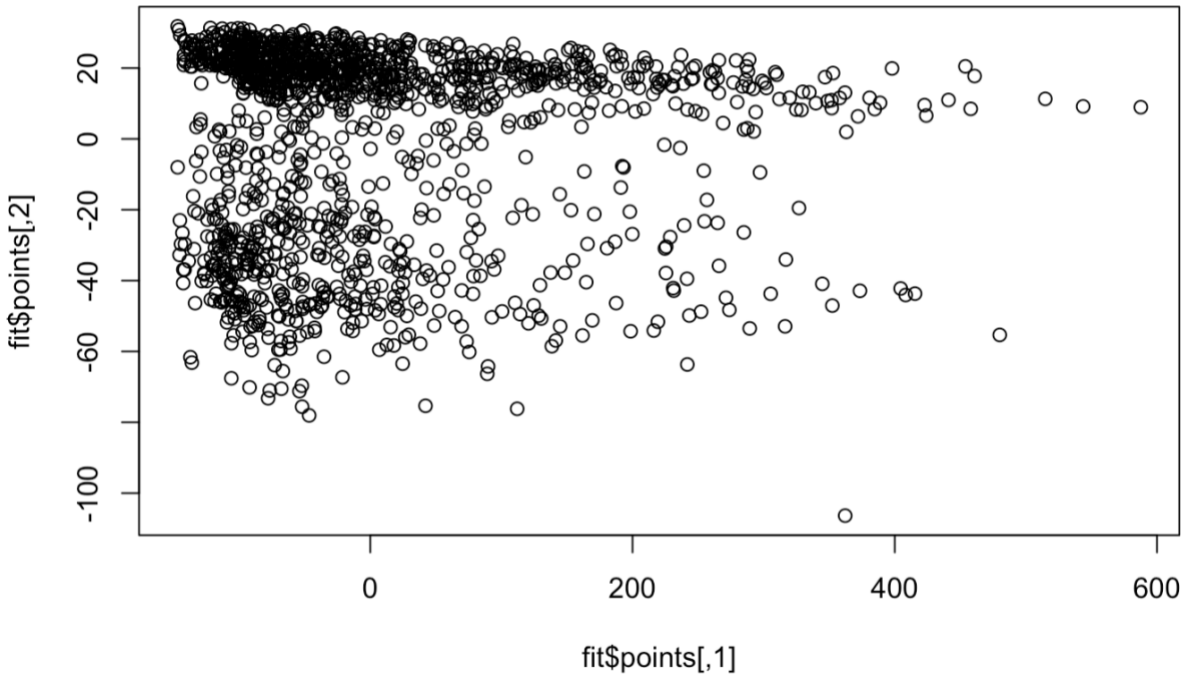
A lot of criminals are in the middle of the education level; this distribution seems quite normal.



Drugs and alcohol seem extremely correlated.

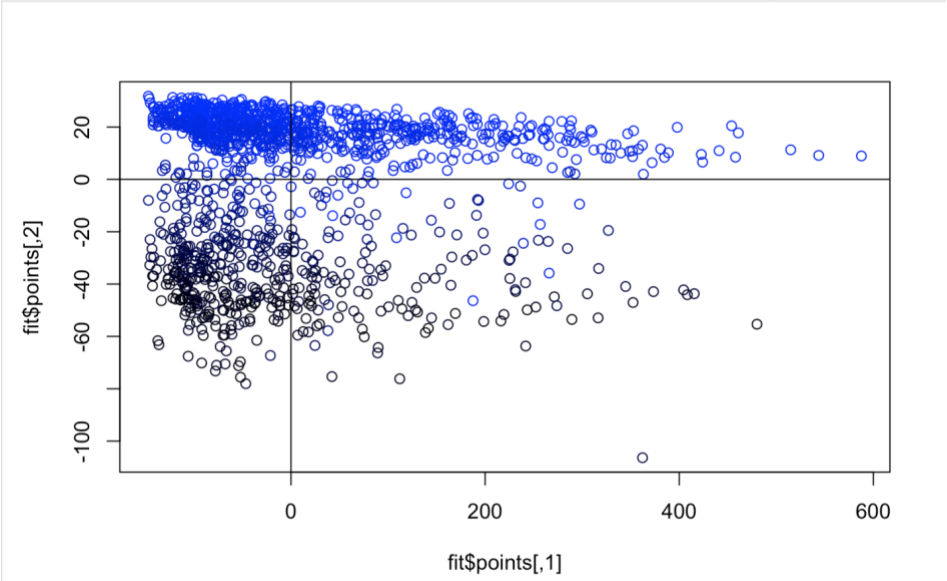
2) Obtain a dissimilarity matrix using Euclidean Distances. There are a lot of cells in this matrix, but can you see any patterns at this point?
There does not seem to be any visible patterns in the dissimilarity matrix.

3) Conduct a classical multidimensional scaling using the Euclidean Distances dissimilarity matrix. Graph a 2-dimensional solution and interpret the result.



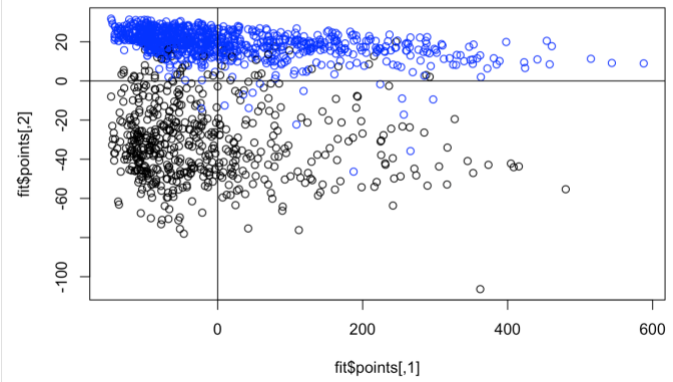
Most data are clustered high in dimension 2 and low/negative on dimension 1.
The most prominent relationship seems to be the separation in duration, as those with high time until they return are mostly in that cluster:

```r
valcol <- (mydata$durat + abs(min(mydata$durat)))/max(mydata$durat  + abs(min(mydata$durat)))
plot(fit$points, col=rgb(0, 0, valcol))
abline(v=0,h=0)
```

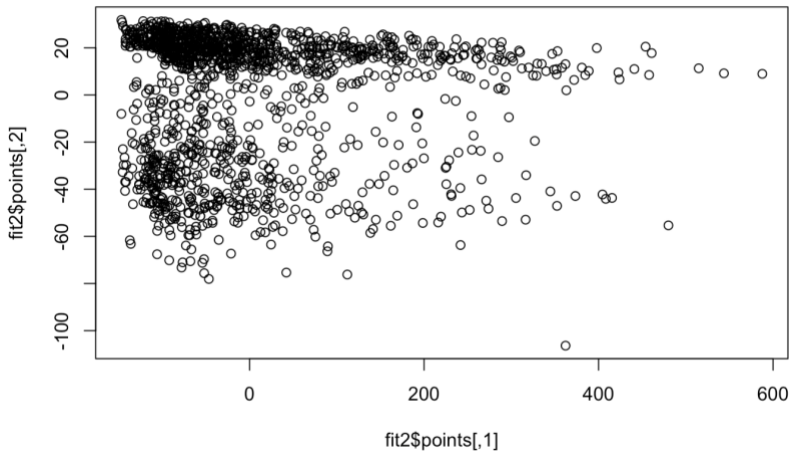


Another good separating variable was the censored variable, who also appeared to be mainly within that upper left cluster:

```r
valcol <- (mydata$cens + abs(min(mydata$cens)))/max(mydata$cens  + abs(min(mydata$cens)))
plot(fit$points, col=rgb(0, 0, valcol))
abline(v=0,h=0)
```

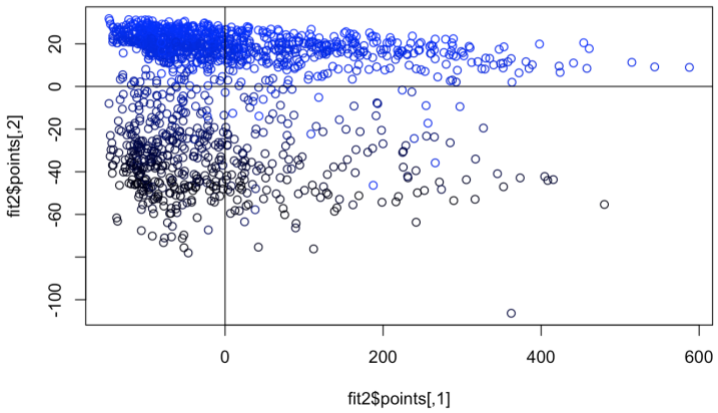


4) Conduct 2 similar analyses using nonmetric scaling and Ramsey's method. Graph and interpret the two-dimensional solutions. How do these solutions compare with the classical approach?
The two methods are very similar to each other. The same relationships as before are visible in this graph as well.



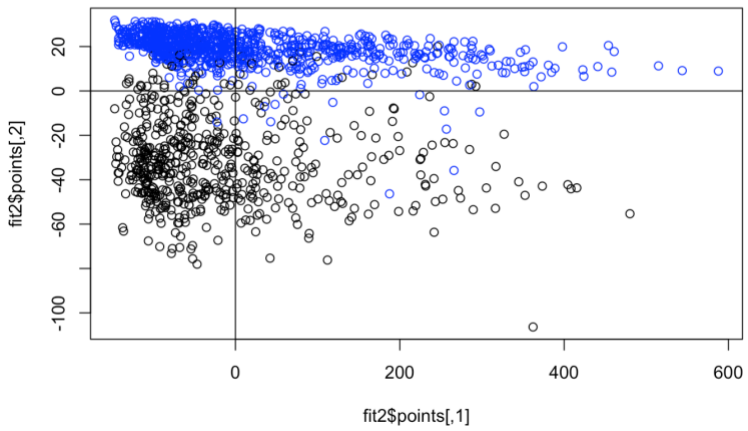
Duration:

```
```{r}
valcol <- (mydata$durat + abs(min(mydata$durat)))/max(mydata$durat  + abs(min(mydata$durat)))
plot(fit2$points, col=rgb(0, 0, valcol))
abline(v=0,h=0)
```
```



Censor:

```r
valcol <- (mydata$cens + abs(min(mydata$cens)))/max(mydata$cens  + abs(min(mydata$cens)))
plot(fit2$points, col=rgb(0, 0, valcol))
abline(v=0,h=0)
```



COMPONENT 2: Self Organizing Maps

In this component of the assignment you will fit Self-organizing Map [SOM] models.

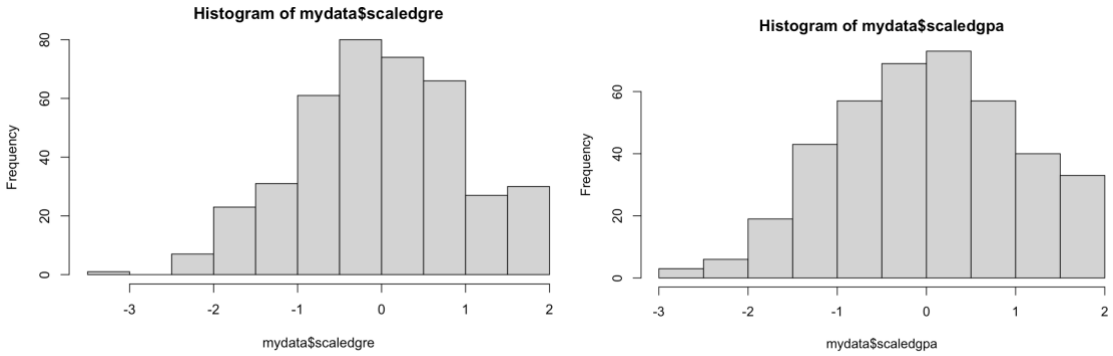
Data: The data for this assignment is the college acceptance data set. The dataset contains information for college acceptance into various engineering programs for 400 students. A simple data dictionary for the dataset is as follows:

- admit (binary): 0 = Not admitted, 1 = Admitted
- gre (numeric): Student's GRE score
- gpa (numeric): Student's GPA
- rank (numeric): College ranking

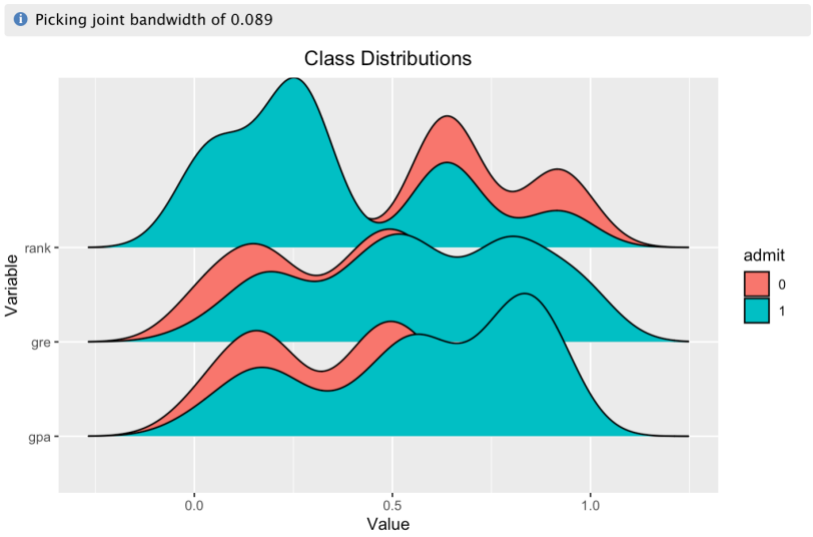
Assignment Tasks

5) Exploratory Data Analysis [EDA] and Data Preparation for the College Acceptance Data.

- Perform EDA on the data set and report your findings.



GPA and GRE scores seem correlated.



The scaled data is more easily comparable.

**On scaled data

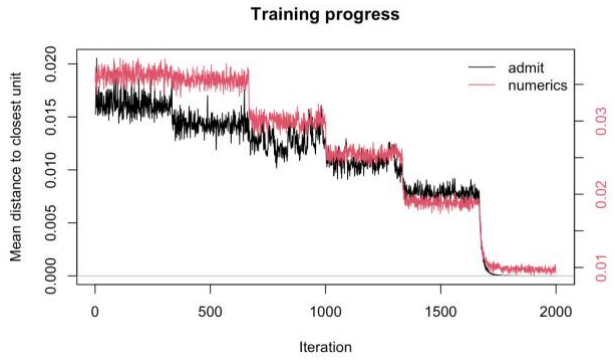
- Prepare the dataset for modeling as appropriate. Should scaling or normalization be applied? Why or why not?
  Yes, as the variables could then be compared on a similar scale, making it easier to analyze trends in the data.
  Also, scaling helps ensure that no individual variable has too much influence on the mapping.
- Use only the variables provided in the dataset or variables you create by modifying or combining the variables provided.
  3 new variables, which are scaled GRE, GPA, and Rank

6) Fit the SOM model. In the process you need to:

- Determine and report the number of epochs that will be used to train the model.
  I used 2000 epochs to train this model.
- Determine the appropriate grid size for the SOM. Report the method that you used.
  I used the find_grid_size method from the classroom page, which gave me a size of 10 to work with for this dataset (meaning 10x10 grid size).
- Fit the model using the R *kohonen* package or similar to the dataset that you prepared in *PART A*. Use the grid size and epochs that you selected in 1 and 2. Be sure to set the seed before fitting the model so that the results may be reproduced.
  Setup:

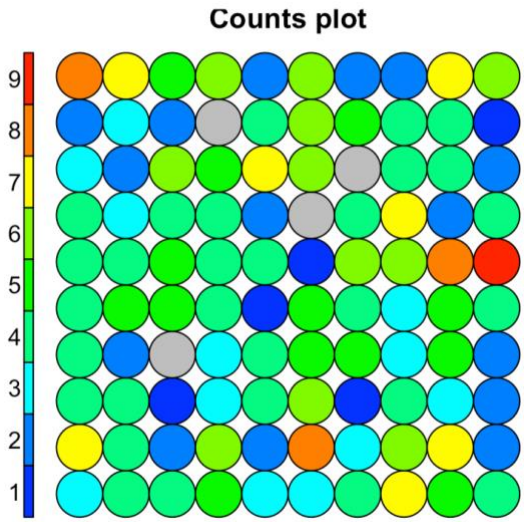
```r
# call the find_grid_size function to calculate the grid size
map_dimension = find_grid_size(dim(raw_data)[1])

# set the number of times the model will cycle through the observations
epochs = 2000
set.seed(123)

# create a grid onto which the som will be mapped
som_grid = somgrid(xdim = map_dimension
                   ,ydim = map_dimension
                   ,topo = "rectangular")

# train the SOM
cc_som = supersom(data_list
                  ,grid = som_grid
                  ,rlen = epochs
                  ,alpha = c(0.1, 0.01)
                  ,whatmap = c(factors, 'numerics')
                  ,dist.fcts = distances
                  ,keep.data = TRUE
)
```

  Results:



7) Evaluate the SOM model. To do this you need to address the following:

- Was the epochs value selected in PART B adequate to train the model? Include a copy of the visualization that was used to make that determination. If the model needs additional training, adjust the epochs value and retrain the model before continuing.
  The epochs value was very good, as the graph shows how close to zero the "Mean distance to closest unit" got to. It got to its lowest point at around 1,700 epochs.
- Was the grid size selected in PART B adequate? Explain why the grid size was or was not adequate and attach the visualizations used to make that determination.



There are four grey or empty nodes, which is not a lot but indicates the size may have been a little more than necessary.

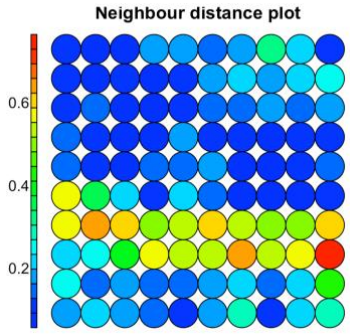
- What is the average number of observations assigned to the nodes?

```{r}
cc_som$unit.classif
observations_by_node <- get_node_counts(cc_som$unit.classif)
```

```
  [1]  95  22  25  11  91  29   7  79  12  34  41  70  25  39  25  85  41  88  36   5  75  17  81  71  28  26  16   2  28  60  94
 [32]  51  74  31  80  79  60  99   9  20 100   8  19  85  39   1   8  91  98  96  43  86  61  18  73  31  75  88  78  81  19  92
 [63]  63  32  54  57  41  60  56  36  43  97  93  55  61  42  74  31  60  15  71  49 100  97   3  59  58  57  37  14  34  26  36
 [94]  50  18  48  83  68  39  96  87  64  86  54  14  29  26  69  88  68  71  86  87  60  52  43   9  46  26  87   4  10  99  76
[125]  52  93  16  41  59  86  19  50  58  76  59  65  92  42  48  16  44  21  94  74  92  96  68  99   7  37  26  78  14  74  53
[156]  11 100   6  47  52  49  47  17 100  58  35  86  52  54  64  96  99  62  28  93  18 100  11  73  87  53  94  34   4  91  36
[187]  64  50  59  68  19  51  53  95   8  67  72   1  64  94  51  19  25  94  15  31  37  16  75  57  61  59  89  73   8  11  80
[218]   6  70   8  75  43   7  51  40  22  57  91  68  28  92  73  78  88  26  49  18  55  91  62  89   5  11  49  60  42  38  79
[249]  63  53  82  82   4   2  21  73   9  13  67  44  39  79   1  13   3  96  92  12  18  99  14  85  17  37  59  19  65  16  11
[280]  19  44  96  83  91   9  92  27  21  61  98  18  40  46  35  90  75  60  69  67  62  48  22   9  14  87  92   7  57  59  76
[311]  54  46  53   2  93  10  10  21   2  60  76  68  91  89  99  35  38   8  58  91  42  63  70  74   3  16  75  82  60  22  86
[342]  20  94  59  85  76  58  55  78  50  24  73  11  47   4  30  38  37  12  42   5   6  39  68  16  99  76  63  45  36   4  32
[373]  30  16  55  93  57  24  72 100  46   8  67  35  10  70  24  58  49  47  29  14  12  23   3  44  75  89  46  43
```

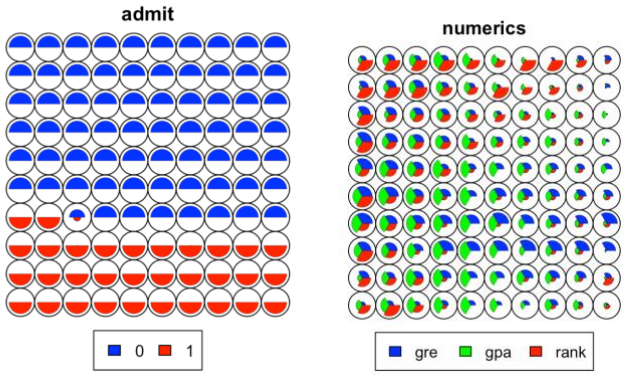
Around 4 (or 4.166667) observations are assigned to each of the nodes on average.

- Generate a distance map and attach a copy of it here. Are any nodes quite distant from their neighbors?



The red dot towards the bottom right seems like a drastic change from those around it, but when we run the code to check which means are drastically different, we don't find any observations.

- Generate a *codes* plot and attach a copy of it. Discuss what this plot tells us about the applications and college acceptance.

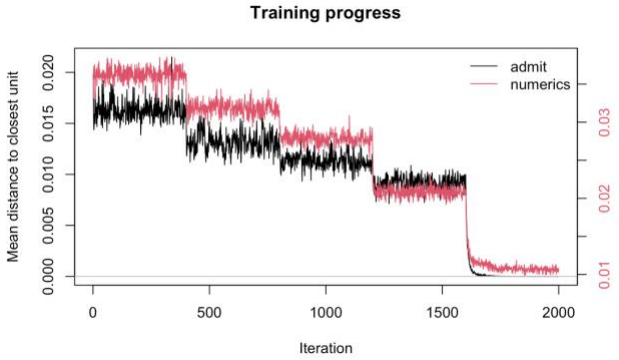


Those with higher GPAs and GRE scores are the ones that got accepted, as a trend is visible with the red appearing at the bottom of the chart. It appears rank has the least deciding factor over acceptance out of the three variables.

8) Experiment with the SOM model. To accomplish this task, you will need to:

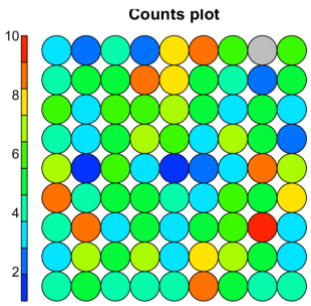
- Change the grid size for the SOM and retrain the model. Discuss whether you increased or decreased the grid size and why.
  I had decreased the grid size to a 9x9 because of some of the grey cells in the previous map.

```
# set the number of times the model will cycle through the observations
epochs = 2000
set.seed(123)

# create a grid onto which the som will be mapped
som_grid = somgrid(xdim = 9
                   ,ydim = 9
                   ,topo = "rectangular")

# train the SOM
cc_som = supersom(data_list
                  ,grid = som_grid
                  ,rlen = epochs
                  ,alpha = c(0.1, 0.01)
                  ,whatmap = c(factors, 'numerics')
                  ,dist.fcts = distances
                  ,keep.data = TRUE
)
...
```

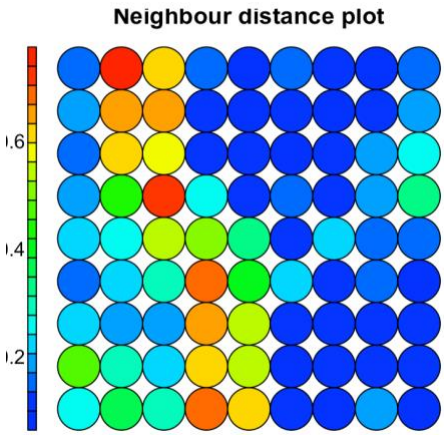


- Compare this new SOM to the SOM created in PART B. Does the new grid size improve the SOM? Discuss how grid size impacts the SOM.

We are now only left with one grey node in the counts plot, which is quite the improvement from last time. The model also zeros out sooner than previous when looking at the admit line in the training process graph. The mean of the observations per node increases to 5.

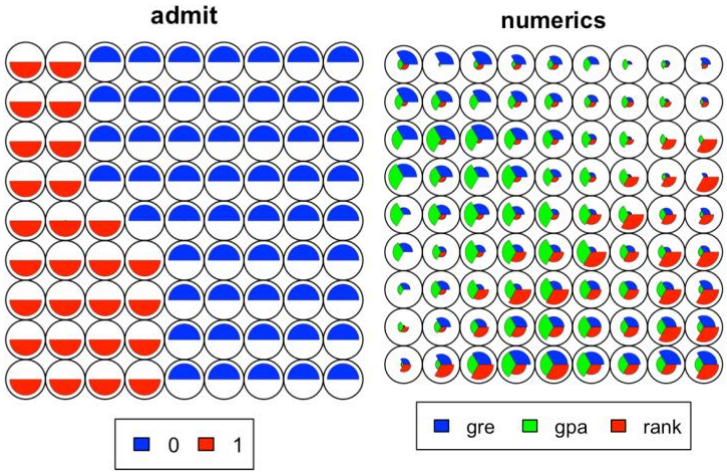


- Generate a distance map and attach a copy of it here. Are any nodes quite distant from their neighbors?



The nodes do not seem as drastically distant as before, however there is still the red node which is now to the center left which raises a bit of concern.

- Generate a *codes* plot and attach a copy of it. Discuss what this plot tells us about the applications and college acceptance.



The model seems to be more confident in its analyses, as the one cell where the model second guessed upon whether the person was admitted or not has disappeared. There is no node where it has both red (1) or zero (0) in the same cell.

9) Please write a reflection on your MDS and SOM modeling experiences.
I believe compared to the previous assignments; this assignment was more straightforward in terms of its expectations. Although daunting at first, once you understand the meaning behind the code and experiments, it is quite easy to grasp the concept. There were lots of methods to figure out what each step was doing, how it worked, and what the meaning behind the steps were.

Assignment Document:

All assignment reports should answer each of the questions separately. Please be sure to clearly indicate which question is being addressed. Results should be presented and discussed in an organized manner with the discussion in close proximity of the results. The report should not contain unnecessary R-code, intermediary computations, R-results, or non-essential information. The document should be submitted in pdf format. Name your file Assign3_LastName.pdf.